Neuropsychiatric Disease and Treatment Dovepress

Open Access Full Text Article ORIGINAL RESEARCH

Computational Psychotherapy System for Mental Health Prediction and Behavior Change with a Conversational Agent

Tine Kolenik[1–3], Günter Schiepek[3,4], Matjaž Gams[1]

[1]Department of Intelligent Systems, Jožef Stefan Institute, Ljubljana, Slovenia; [2]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia; [3]Institute of Synergetics and Psychotherapy Research, University Hospital of Psychiatry, Psychotherapy and Psychosomatics, Paracelsus Medical University, Salzburg, Austria; [4]Department of Psychology, Ludwig Maximilian University of Munich, Munich, Germany

Correspondence: Tine Kolenik, Email tine.kolenik@ccsys.de

Background: The importance of computational psychotherapy is increasing due to the record high prevalence of mental health issues worldwide. Despite advancements, current computational psychotherapy systems lack advanced prediction and behavior change mechanisms using conversational agents.

Purpose: This work presents a computational psychotherapy system for mental health prediction and behavior change using a conversational agent. It makes two major contributions. First, we introduce a novel, golden standard dataset, comprising panel data with 1495 instances of quantitative stress, anxiety, and depression (SAD) symptom scores from diagnostic-level questionnaires and qualitative daily diary entries. Second, we present the computational psychotherapy system itself.

Hypothesis: We hypothesize that simulating a theory of mind - the human cognitive ability to understand others – in a conversational agent enhances its effectiveness in relieving mental health issues.

Methods: The system simulates theory of mind with a cognitive architecture comprising an ensemble of computational models, using cognitive modelling and machine learning models trained on the novel dataset, and novel ontologies. The system was evaluated through a computational experiment on mental health phenomena prediction from text, and an empirical interventional study on relieving mental health issues in 42 participants.

Results: The system outperformed state-of-the-art systems in terms of the number of detected categories and detection accuracy (highest accuracy: 91.41% using k-nearest neighbors (kNN); highest accuracy of other systems: 84% using long-short term memory network (LSTM)). The highest accuracy for 7-day forecasting was 87.68%, whereas the other systems were not able to forecast trends. In the study, the system outperformed Woebot, the current state-of-the-art, in reducing stress ($p = 0.004$) and anxiety ($p = 0.008$) levels.

Conclusion: The confirmation of our hypothesis indicates that incorporating theory of mind simulation in conversational agents significantly enhances their efficacy in computational psychotherapy, offering a promising advancement for mental health interventions and support compared to current state-of-the-art systems.

Keywords: generative artificial intelligence, attitude and behavior change support systems, digital mental health, intelligent cognitive conversational agent, artificial cognitive architecture, machine learning

Introduction

Currently, the world is experiencing a mental health pandemic.[1] There are numerous calls for action, even to the degree to which mental health is now included in Goal 3 of the 17 UN Sustainable Development Goals.[2] However, the number of mental health professionals per 100.000 people is declining, not rising, and current regulations do not seem to improve the situation.[3] Stress, anxiety, and depression (SAD) are at the forefront of mental health issues; in some groups, figures reach 74% for disabling stress,[4] 28% for anxiety disorder[5] and 48% for depression.[6] Between 76% and 85% of people in low- and middle-income countries receive no treatment for their issues.[7] In high-income countries, treatment coverage for depression was only 33%.[8] Many countries have struggled with high suicide rate.[9] The COVID-19 pandemic has

Received: 18 June 2024
Accepted: 14 November 2024
Published: 12 December 2024

further exposed how harmful neglecting people's well-being for decades can be,[10] with social distancing helping in the exponential rise in psychopathological symptoms.

Mental health issues have considerable effects on patients, their immediate surroundings (family, caretakers), and wider society.[11] This may be why decision-makers are turning to technology to see how it can contribute toward easing the load on mental healthcare. A recent analysis[12] identified the possible positive and negative effects that using technology might have in mental healthcare, where the former seems to outweigh the latter. Technology can lower costs for patients, make help available at all times (eg, during a nighttime panic attack) and in more places (eg, remote places with no health facilities nearby), reduce stigma (making help available to people who are too afraid to visit a professional), and boost prevention of mental health issues (professionals commonly enter the treatment process only when mental health issues already exist, whereas technology can help prevent issues from appearing). However, some groups may be excluded from using technology (eg, the elderly), and there is a higher probability of research bias, since digital mental health is a new research field with fewer standards. Regardless, technologies entering mental healthcare seem to have reached certain populations that were previously left out.[13]

Introducing novel ways to include therapeutic approaches in people's lives, especially through technology, might also help destigmatize psychotherapy. 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] was recognized. Such technology can therefore be placed under the umbrella of computational psychotherapy. This encompasses the following: (1) studying psychotherapeutic processes computationally, using client and professional data to create various analyses and models and (2) creating computational tools for mental health care and supporting existing mental health professionals, which in certain cases take the role of artificial intelligent support, making use of psychotherapeutic approaches and techniques such as cognitive behavioral therapy. Persuasive technology (also called attitude and behavior change systems) is an example of a technology that can be used in computational psychotherapy.

Persuasive technology (PT) represents a synergy between progress in behavioral sciences and many recent advances in computer science, especially artificial intelligence (AI), which is increasingly characterizing our information society.[15] Its goal is to "change attitudes or behaviors or both (without using coercion or deception)"[16] and to "aid and motivate people to adopt behaviors that are beneficial to them and their community while avoiding harmful ones".[17] An effective technological vessel for persuasion is an intelligent conversational agent (also known as a chatbot or an 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 (CogArch), 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]

Most importantly, ICAs possess the ability to converse in natural language, which is likely the most immediate way in which humans communicate,[20] and interacting through dialog is extremely important in the field of computational psychotherapy and digital mental health.

However, it is challenging to develop effective technologies for mental health support. Even mental health professionals and practitioners find it exceedingly difficult to detect idiographic specificities of a person's mental health status and intervene effectively. The mental health status of a person is very dynamic, meaning that it changes nonlinearly over short time ranges. Experiments seem to indicate that computational psychotherapeutic systems can successfully mitigate symptoms of SAD.[21–25] 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,[26] and the most advanced language models, such as recent GPTs,[27] have a history of underperforming domain-specific tasks, going even as far as telling depressed patients to kill themselves.[28] When technology is used to help people with SAD issues, it can underperform because of the use of nomothetic data and non-personalized interventions. This is underpinned by the fact that no datasets are available for

constructing truly successful systems in the first place. One of the reasons is that collecting quality data in this domain is challenging. However, ecological momentary sampling (EMA) is proving as an effective framework for quality data collection. This method follows the "anywhere, anytime" (A–A) principle,[26] using a smartphone to collect data in non-invasive ways. EMA transports data collection from the laboratory into the real-life setting. This has several benefits, as people cognize differently in the wild, in their ecological environments, than in the lab,[29] 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.[30]

Thus, the participants are in their ecological environment when the data are collected. The data collection can be signal-contingent, which "constitutes randomising notification timing throughout of a given timespan",[31] interval-contingent, in which "notification timing is scheduled in line with a (predefined) time gap",[31] eg, a questionnaire every 2 hours, or event-contingent, where "the occurrence of a predefined event results in a notification",[31] eg, the smartphone senses specific movement through the global positioning systems (GPS) data.[32] EMA is being successfully used in mental health research.[33]

Therefore, a holistic and integrative approach to creating a system with novel idiographic models based on human cognitive capabilities is required to overcome these issues. This work presents such a system, as well as a novel dataset that drives it. The authors focused 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 into the mental healthcare system, but could also be used complementarily by mental health professionals, eg, 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 CogArch that uses not only AI and machine learning (ML) but also other novel mechanisms. It simulates the theory of mind, the human cognitive ability to understand people, and how to effectively respond to them. This is achieved by building various idiographic, detection, and forecasting models in novel ways that enable them to perform with higher accuracy than the current SOTA. These models are combined with novel ontologies on mental health and behavior change, and are used to understand ICA users. A precise and careful research design was used to collect ecological and viable time-series data from individuals with SAD symptoms. The methods used strove toward explainable and open AI. Furthermore, existing large language models (eg, GPTs) have seen little specialization for complex domains such as mental health, and this CogArch accommodates these models through novel integration into its architecture and makes them usable to help people with SAD.

The following sections provide a detailed overview of the work. The Related Work section, which follows, reviews existing systems in the field of intelligent cognitive assistants for mental health, highlighting their capabilities and limitations. The Rationale for this work outlines the motivation behind developing this novel system. Research Goals and Hypotheses defines the key objectives of the study and the main hypothesis. The Materials and Methods section describes the dataset, machine learning models, and experimental designs used. Cognitive Architecture Design details the system's architecture, including its modules and functions. The Results section presents the findings from computational experiments and the empirical interventional study. The Discussion compares the system's performance with state-of-the-art solutions, its strengths and limitations, and potential avenues for future research. Finally, the Conclusion summarizes the contributions and offer concluding thoughts.

Related Work

The authors recently published a substantial SOTA technical review paper on ICAs for attitude and behavior change support in mental health.[35] The paper focuses on non-proprietary systems. The authors also explored SOTA proprietary systems in a recent paper,[36] although in less detail, as proprietary systems are harder to dissect technically because they are not open code. There are also no papers discussing their technological underpinnings, particularly through computational experiments. A book chapter by the first author offers a detailed review of ML and similar methods used in SOTA in the field of mental health.[26] This section briefly summarizes these works.

Non-Proprietary Systems

The review of non-proprietary systems followed Arksey and O'Malley's framework[37] and PICOC methodology[38] and is detailed in the previous authors' work.[35] These systems present various approaches to achieving change in people with mental health issues.

Short Description of SOTA Systems

Below is a short summary of each reviewed SOTA system, highlighting its distinctive features:

• Delahunty et al[39] proposed a diagnostic ICA, which combined conversational abilities with ML and clinical psychology to facilitate crisis support for people with depression. It used sequence-to-sequence neural networks for dialog generation and machine learning classifiers to discover depression symptoms.

• Denecke et al[40] introduced SERMO, an ICA that combined methods from cognitive behavior therapy (CBT) and lexicon-based emotion recognition to support the general well-being of people by regulating their emotions, thoughts, and feelings. Additionally, informational strategies have helped provide people with psychoeducation.

• Ghandeharioun et al[41] concentrated on providing ecological momentary interventions (EMI) via an ICA to enhance an individual's overall well-being by alleviating SAD symptoms. The EMMA system offered emotionally suitable interventions empathetically, identifying users' moods exclusively through smartphone sensor data.

• Khadikar et al[42] created Buddy, an ICA aimed at improving general well-being by addressing SAD symptoms and serving as a motivational companion to assist with lack of focus. The system utilized recurrent neural networks (RNNs) to generate fitting dialogs in response to users' emotions.

• Morris et al[43] developed an ICA that emulated human empathetic expression abilities. They repurposed online peer support data, which the ICA presented to users by utilizing information retrieval and word embedding techniques.

• Park et al[44] introduced a prototype ICA called Bonobot that employed motivational interviewing techniques to help students manage their stress. The ICA guided users through motivational interviewing processes using conversational sequences, offering provocative questions, supportive feedback, as well as reflective and affirmative responses within the context of users' issues.

• Pola and Chetty[45] designed an ICA that delivered behavioral therapy to individuals with depression. The ICA sought information on users' mental states and was able to identify seven types of emotions from text using long-short-term-memory neural networks and a pre-trained weighted word index called glove2.

• Rishabh and Anuradha[46] developed three distinct ICAs for general well-being, each utilizing various technologies. The first, based on ELIZA,[47] employed retrieval methods for language abilities. The second, inspired by ALICE,[48] used AIML (Artificial Intelligence Markup Language). The third adopted generative techniques. All three attempted to understand the context users conveyed through text and to steer the conversation toward a more positive sentiment.

• Yorita et al[49] suggested a stress management framework that incorporated an ICA platform. The ICA calculated multiple-stress metrics and modeled users, which informed the strategy selection in their peer support model. The system aimed to teach users different stress management techniques, driven by a combination of reinforcement learning and fuzzy control.

• Yorita et al[50] expanded upon the ICA from Yorita et al,[49] further refining the models and strategies for assistance to personalize their system even more.

The following six sections (taken from[35] with the authors' permission) are as follows.

User Assessment in SOTA Systems

This section provides an overview of the methods and models used by the reviewed systems to infer the mental states of users and guide support, as outlined below:

• Delahunty et al:[39] The system identified depression, suicidal thoughts, insomnia, hypersomnia, weight fluctuations, and excessive or misplaced guilt from users' linguistic input. It was trained on multiple datasets, including eRisk and Reddit posts from users and specific subreddits. The system utilized doc2vec for feature extraction and employed Random Forest and Logistic Regression to predict the presence or absence of depression symptoms, achieving an F1-Score of 0.91.

• Denecke et al:[40] The system leveraged the SentiWS lexicon for sentiment and emotion detection in text input using fuzzy matching techniques. It attained 81% accuracy in identifying emotions in a dataset comprised of forum posts.

• Ghandeharioun et al:[41] The system gathered geolocation information from a phone and implemented experience sampling five times daily using a visual grid based on Russel's two-dimensional emotion model to obtain ground-truth labels. The system achieved 82.4% accuracy in valence prediction using Random Forest and 65.7% accuracy in arousal prediction using AdaBoost Regression.

• Khadikar et al:[42] The system did not explicitly assess users. Assessment was implicit in the linguistic intent recognition in the conversational model.

• Morris et al:[43] The system did not explicitly assess users. Assessment was implicit in the linguistic intent recognition in the conversational model.

• Park et al:[44] The system used evocative questions to collect linguistic user input. Subsequently, it used keywords from the linguistic user input to guide the conversation – these keywords conveyed mental states. The keywords were acquired from a dataset that collected data from the Reddit subreddits.

• Pola and Chetty:[45] The system used questions that target emotional states of the users. It then used a model to detect seven types of emotions. The model used long-short-term-memory (LSTM) neural network with glove2 for emotion recognition. The model was trained on the ISEAR dataset. The accuracy of the emotion recognition obtained was 84%.

• Rishabh and Anuradha:[46] The system did not explicitly assess users. Assessment was implicit in the linguistic intent recognition in the conversational model.

• Yorita et al:[49,50] The system used fuzzy inference to evaluate the content of the linguistic user input as replied to various intentional questions to detect users' state of stress. This was used for various strategies to increase stress management.

Interventions and Effectiveness in SOTA Systems

This section provides an overview of the methods the reviewed systems used to intervene with users and offer mental health support, along with an evaluation of their interventions and performance, as outlined below:

• Delahunty et al:[39] The system did not provide interventions and lacked specific intervention techniques. Its effectiveness was not evaluated.

• Denecke et al:[40] The system's conversational model was constructed using the Syn.Bot framework, which employs Oscova as the bot development platform and the SIML (Synthetic Intelligence Markup Language) interpreter. The model allowed users to phrase answers in their own words and chose predefined responses. The system offered suggestions for activities and exercises to regulate emotions through dialog, reminded users of appointments, and implemented CBT techniques such as mindfulness and goal focusing. The dialogs varied based on the detected emotions and were primarily informational. Both users and mental health professionals tested the system on 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).

• Ghandeharioun et al:[41] The system's conversational model relied on textual prompts and pre-scripted phrases employed at contextually suitable moments.

The system provided well-being interventions that encompassed individual or social activities from various psychotherapeutic categories: positive psychology, cognitive-behavioral, meta-cognitive, or somatic interventions. These were delivered to the user through a textual prompt, accompanied by different digital tools to engage in the activity. The system generated dialog based on the detected emotions by randomly selecting a pre-written script from a corresponding emotional category.

The system did not provide interventions and lacked specific intervention techniques.

• Khadikar et al:[42] The system's conversational model used RNNs for learning as well as understanding and generating responses. The intent in the user input was recognized by the LSTM neural network.

The system delivered interventions in the form of positive drivers inserted into the conversation to change the trends of users' thoughts. It also targeted self-expression development and stress management. CBT techniques, motivational

interviewing and analysis, positive behavior support, behavioral reinforcement, and guided actions and methods were used to develop emotional resilience skills. The system did not deliver any interventions and had no specific intervention methods.

There was no evaluation on its effectiveness.

• Morris et al:[43] The system's conversational model was composed of two modules. The front-end module connected previous responses to user inputs, whereas the back-end module generated output using Elasticsearch, word2vec, and a word-embedding process. The authors utilized the Google News dataset for training purposes.

The system offered interventions in the form of pre-existing emotional support statements, sourced from a large corpus of online interactions on the Koko platform, which connects users seeking help with those willing to provide assistance. Users in need of help also assessed the responses. This corpus-based approach aimed to create an impression of personalized, empathic expression. The system employed information retrieval techniques and word embeddings to automate this process in real-time, matching existing statements with suitable user inputs and selecting texts with adequate scores.

Users compared the system's responses to their peers' responses using three ratings: *good* (system: >40%; peers: >60%), *ok* (system: <40%; peers: <40%), and *bad* (system: >20%; peers: <10%).

• Park et al:[44] The system provided interventions in the form of motivational interviewing, relying solely on predefined responses that varied depending on the user's progress stage. These stages, as defined in the motivational interviewing method, include Engaging, Focusing, Evoking, and Planning. This process aided users in coping with stress and promoting self-reflection.

Users described the system as having thought-provoking questions and encouraging self-reflection and potential consolidation, but they observed that the feedback was clichéd. Users also desired more informational support and better contextualized feedback from the system.

• Pola and Chetty:[45] The system offered interventions as emotional conversational support, suggesting alternative, more positive outlooks on the situations users described, and attempting to prevent negative thoughts. The conversation was guided by the severity of the detected mental health issue.

The system's effectiveness was not evaluated.

• Rishabh and Anuradha:[46] The systems employed various approaches for delivering interventions. The ELIZA-based ICA utilized Rogerian reflection to interact with users, employing information retrieval techniques such as the n-gram technique, charagram embeddings, word similarity, sentence similarity, and part-of-speech tagging to choose appropriate responses. The ALICE-based ICA provided interventions by empathizing with the user and applying CBT techniques, using AIML, sklearn for matching responses, and category tagging and synonym switching for conversational dynamism. The generative language ICA implicitly delivered interventions by training on empathetic dialogs.

The systems' effectiveness was not evaluated.

• Yorita et al:[49,50] The system provides interventions that enhance the users' stress management. It employs reinforcement learning and fuzzy logic to optimally match Peer Support strategies (helper therapy, informational, esteem, and emotional support) with specific Sense of Coherence user profiles.

The system was evaluated using daily questionnaires during usage to discover a statistically significant increase in users' stress management scores by 20%.

Proprietary Systems

Two proprietary ICA systems are described in this section.

Tess "reduce[s] self-identified symptoms of depression and anxiety".[51] It uses an extensive ontology on emotions. This ontology is used on the input text to discern users' mood. After mood assessment, Tess uses scripts to help the user. Once Tess dispatches the help, it gathers journaling data and user feedback to improve them. When tested, depression and anxiety symptoms in the test group, which used Tess, were reduced by roughly 15%, whereas the control group, which used official self-help material, saw no change.

Woebot primarily functions using a

decision tree with suggested responses that accepts 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.[52]

Its user model collects data on users' moods, goals, expectations, and similar. The user model guides Woebot's intervention selection, which can be in the form of educational videos and tailored advice. When used in a randomized controlled trial, the test group, which used Woebot, saw 20% SAD symptom relief, whereas the control group, which used the government-approved self-help book, saw no change.

Woebot currently appears to be the best performing available ICA for mental health support, according to conducted research, with most studies showing positive effectiveness.[52,53]

Rationale for This Work

To further advance the SOTA research presented in the previous section, this work attempts to interdisciplinarily combine AI, attitude and behavior change, and mental health in a novel synthesis.

The main idea is to create a computational simulation of theory of mind, a human cognitive capability. Theory of mind is at the core of this work's system. In many soft domains, it is theory of mind that distinguishes humans from machines. This can enable using AI for a very specific, mentally dynamic and complex task – attitude and behavior change in mental health.

The relevance of this work for AI lies in building and combining real-time machine learning-based detection models, where we aim for SOTA accuracy; forecasting models, which do not exist yet in this way in mental health; novel ontologies on mental health and behavior change; and recent large language models, which did not succeed in mental health in the past.[28] Wrapping large language models to generate language in prompt generators, which rely on cognitive and personality models of users with the use of behavior change theories, as well as including risk-filtering models, makes sensible motivational message generation possible for SAD symptoms relief.

This is enabled through the creation of a novel, so far nonexistent 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.[54] This explains the intersection between AI and mental health.

Finally, focusing on attitude and behavior change, recent experimental work in behavioral sciences showed that knowing certain personality types of people makes them more susceptible for influence through specific persuasive strategies.[17] This work uses these advances and adapts them for mental health through the use of technology, which makes the personalization of strategies very viable. So far, personalization in behavior change has mostly been used in static environments,[55] where personalization is harder, which makes strategies less effective and specific.

Research Goals and Hypotheses

Research Goals

The main goal of this work was to design and implement a novel artificial cognitive architecture that will mimic theory of mind, which is in cognitive science described as the cognitive ability to "understand the thoughts and feelings"[56] 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 decision trees, random forest, and various neural network architectures, including deep neural networks. Recent large language models[57,] have been included as well. They

generated linguistic outputs as a response to the input text where 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. Because recent large language models are not adapted to specific domains as well as prone to risky output,[28] 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 new strategies if the well-being had been improved for future use.

Another goal was to produce a novel mixed methods dataset with more than 1000 data instances that was 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 and a main study (the latter ethically approved). This ensured the golden standard data needed for this work's system as well as a dataset that can be used by the wider research community that currently lacks such data.

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 Section Related Work);

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,[52] 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.[52]

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;[58]
- a fully developed front-end and user interface, which tends to keep users' attention and focus longer;[59]
- the ability to deliver visual outputs (eg, 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 the theory of mind can enhance the performance of the system developed and described in this thesis.

Materials and Methods

This Section describes the methods and materials used that were necessary to assemble the system described in the previous Sections. 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, feature importance methods (used for novel insights into mental health phenomena), and accuracy measures;
4. Empirical interventional study research design and description.

Data Collection

Research points toward the quality of data being the primary determinant of a successful ML model or an AI system.[60,61] This means that the data used has a greater impact on the model's or system's performance than selecting and building the optimal algorithm. Because 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 a priority 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 four 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 were collected using Google Forms and the Synergetic Navigation System (SNS) application.[62]

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

The ecological momentary assessment (EMA) was employed to collect quality data. This work also utilized an interval-contingent data collection, using Google Forms and SNS[62] 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.

Data Collection Pre-Study

The pre-study was conducted with 8 participants recruited through convenience sampling and lasted for approximately 3 weeks, taking place in June and July 2021. The participants received instructions on how to use SNS and how the data were collected. The data were collected at the end of each day using the 10-item Positive and Negative Affect Schedule (PANAS) questionnaire,[61] and a textual diary entry with specific guidelines (see *Supplementary Material: Pre-study Diary Entry Guidelines*).

Furthermore, the participants completed a demographic questionnaire (see *Supplementary Material: Study Demographic Questionnaire*) and the BFI-10, a 10-item scale measuring the Big Five personality traits Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness.[63] The latter especially served to personalize the linguistic output of the system (for more details, see Section Cognitive Architecture Design).

PANAS, due to its simplicity, was used to understand how to maintain 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.

Furthermore, feedback from participants on their experience using data collection tools (the SNS tool), their thoughts on how the questions, instructions, and guidelines were collected, as well as reports on any other issues or ideas. Before the pre-study began, the participants signed informed consent forms. No reward was provided to the participants for their involvement in the study.

The collected data were not shared with anyone, were held on a secure server, and were used only for research purposes. If at any point the participants decided to withdraw their data, it was deleted from the server. The data were completely anonymized following the protocol of Olden et al for epidemiological or clinical studies.[64]

The takeaways from the pre-study were the following:

- Reformulate text to be clearer;
- PANAS was confusing for participants, it is 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 (important for later ML purposes);
- the time to complete a daily sample was around 15 minutes;
- the questionnaire could be longer in the main study;
- 150 words for the diary was the right amount;
- SNS is appropriate for data collection.

Main Data Collection Study

The main data collection study, as the pre-study, used SNS to collect the data using EMA. The participants were recruited through convenience sampling. The study took place in August and September 2021. The data were anonymized using the protocol described by Olden et al for epidemiological or clinical studies.[64] 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,[65] Beck Anxiety Inventory,[66] Beck Depression Inventory,[67] and Ratcliffe's Depression Questionnaire[68] were used to compile the questionnaire. The final 18-item questionnaire, consisting of 18 questions relating to SAD symptoms, are available in *Supplementary Material: Study 18-Item SAD Questionnaire*.
2. Sixty-one 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 revised guidelines for the diary entry are available in *Supplementary Material: Study Diary Entry Guidelines*.

Of the 61 participants who applied, seven participants never started the study. As in the pre-study, the participants completed a demographic questionnaire, BFI-10,[62] and a post-study questionnaire to ensure the quality of the collected data.

Data Quality

The post-study questionnaire, available in *Supplementary Material: Post-Study Questionnaire*, 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 authors.
- The majority found 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 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.[69]
- None of the participants replied that they were unwilling to participate in a similar study again. This ensures that the data collected were from willing and engaged participants, which also contributed to the low attrition rate in the study.

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

The study was reviewed and approved by the ethics committee of the Collegium of the Department of Intelligent Systems, Jožef Stefan Institute (*Ethical approval code: cafiancimhumema_2021-07-13*), and it was carried out in accordance with the ethical principles in the Declaration of Helsinki. The participants signed informed consent forms before participating in the study. No reward was provided to the participants for their involvement in the study. The full instructions, available to the participants, are available in *Supplementary Material: Study Instructions*.

Dataset

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

One data instance (row) consists of the attributes seen in Table 1.

The dataset consists of 1495 data instances, of which 1168 instances are without missing data. Table 2 represents general statistics about the dataset and its number of instances. The general descriptive statistics of the various attributes in the dataset are available in *Supplementary Material: General Dataset Statistics*, represented in Tables S1–S12, covering statistics on the number of instances and diary word count (Table S1), demography (Tables S2–S11), and Big Five Personality Traits (Table S12).

Computational Experiments

This section describes the methods used for computational experiments performed. The methods were selected according to the overviewed SOTA in Section Related Work. The results are presented in Section Results. The 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 Table 3). Tasks are the following:

1. Detection of SAD levels and SAD symptoms from text. Training occurs on the described dataset, whereas detection occurs only from one text input from a user.

Table 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 Material: Study 18-Item SAD Questionnaire for attributes	Time series	18
Daily diary entry	/	Time series	1

Notes: Attributes in one data instance are summarized across metadata, demographic, personality, daily mental health, and daily diary entry categories, detailing their type (cross-sectional or time series) and the number of attributes per category.

Table 2 General statistics about the dataset and its number of instances.

Per Person	*M*	*SD*
Instances	26.55	7.77
Diary word count	188.81	148.83

Notes: Basic statistics (mean (*M*) and standard deviation (*SD*)) on the number of instances and diary word count per person.

Table 3 Description of SAD levels and SAD symptoms, derived from the daily quantitative questionnaires.

Target Variable	Description
Stress	Sum of Q1, Q2, Q1
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	Sum of Q9–14, Q17–18

Notes: Target variables used in the system's ML (machine learning) models to detect or forecast from the users' text input.

2. Forecast of SAD levels and SAD symptoms from only one text diary entry. Training occurs on the described dataset, whereas the forecast occurs only from one user text input (and not a time-series data input, which are common requirements of forecasting models).

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

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, process of feature engineering, and derived target variables.

Machine Learning Algorithms

Machine learning (ML) algorithms are capable of constructing models that gain experience from collected data. Usually, the more experience that a model has, with the more data that are collected, this improves accuracy in classification of a new data instance.

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

- Decision Tree[70]
- Bagging Decision Tree[70]
- Boosting Decision Tree[70]
- Random Forest[70]
- Complement Naive Bayes[71]

- K-nearest Neighbors Classifier[72]
- Multiple Layer Perceptron Classifier[73] (see Figure S1 in Supplementary Material: ML Algorithms Used).
- Logistic Regression[70]
- Support Vector Machine[74]

The algorithms' technical descriptions can be found in *Supplementary Material: ML Algorithms Used*.

Feature Selection

Two methods were used for the feature selection: Granger causality and collinearity. The description of both, along with relevant formulas, can be found in *Supplementary Material: Feature Selection*.

Feature Engineering

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

Below are presented feature engineering frameworks used in this work.

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.[75] The lexicon detects two sentiment dimensions: polarity and intensity.

Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count (LIWC) is a framework based on lexicon techniques.[76] It forms word categories based on the psychological meaning of words. It enables positioning instances into a multi-categorical space. Among others, LIWC lists the following categories (with relevant Tables of what these categories encompass alongside dictionary examples in *Supplementary Material: LIWC Categories Tables*: Standard linguistic dimensions (Table S13), Psychological processes (Table S14), Personal concerns (Table S15), and Spoken categories (Table S16).

With its use, text can be classified into a multi-dimensional space, providing insights into a variety of mental processes:[76] the topic covered, the style of thinking, the emotional state, the cognitive processes, etc.

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 combined SAD symptoms questionnaire described in Section **Main Data Collection Study** (also see *Supplementary Material: Study 18-Item SAD Questionnaire*). Table 3 describes these variables, composed of mental health issues (SAD) and symptoms (others).

All target variables were binarized where the positive class refers to the target variable being significantly present in the users' text. A binarization threshold formula was used:

$$b = 0.25 \times t$$

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

Categorization of a mental health issue or a symptom that is significantly present at 1/4 of the maximum value of the questionnaire scale can be found in numerous mental health diagnostic tools for SAD.[64,77]

Empirical Interventional Study

To test the efficiency and success rate of this work's system, an empirical interventional study[78] was designed. It compared a state-of-the-art chatbot for attitude and behavior change in mental health, Woebot (for its overview, see Section Proprietary Systems), with this work's system. 42 participants were recruited through convenience sampling, and a short screening questionnaire assessed their demographic and mental health status. The study took place in June and

July 2022. 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,[79] 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").

2. Quantitatively evaluating the experience with chatbots with two expert measures from the User Experience Questionnaire (UEQ).[80] 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 and the *test* group. 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 are involved. This was the study's exclusionary criteria. The data were fully anonymized with the researchers disposing of the data, not present in the final dataset (eg, e-mail addresses), within 1 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. No reward was provided to the participants for their involvement in the study. The study was reviewed and approved by the ethics committee of the

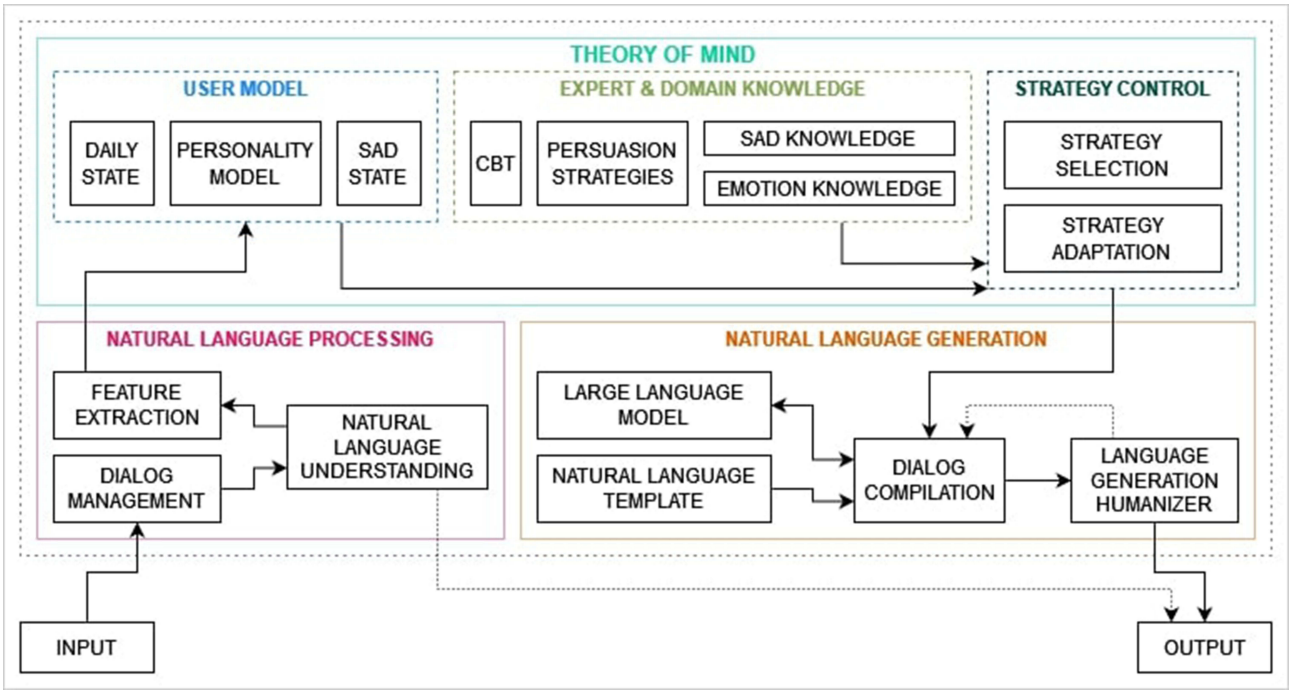

Figure 1 The system's cognitive architecture. It consists of: The *Natural language processing module* (rose color), the *Natural language generation module* (Orange color), and the *Theory of mind module* (teal color), which is further divided into the User model (blue color), the *Expert domain knowledge module* (light green color), and the *Strategy control module* (dark green).

Collegium of the Department of Intelligent Systems, Jožef Stefan Institute (*Ethical approval code: cbssotacfaabcfmh-wasdcips_2022-06-29*), and it was carried out in accordance with the ethical principles in the Declaration of Helsinki.

Cognitive Architecture Design

This section describes the cognitive architecture of our proposed system. The design is illustrated in Figure 1.

The next sections describe the CogA's modules in detail, and describe how the modules make the system work, providing examples, and underlying how each CogA module's performance is enabled by a computational method (See Section Computational Experiments 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).

Natural Language Processing

The Natural language processing module (rose color in Figure 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 tracks 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.

Dialog Management

The Dialog Management submodule keeps track of where the ICA and the user are in the conversation. 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),[81] as the dialog tree consists of data-modelling as well as probabilistic sequencing depending on the ML detection of user states.

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[82] to identify invalid linguistic inputs from a predefined 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.

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. The main methods used include the LIWC framework[76] and the VADER sentiment model.[75] See Section Feature Engineering for the in-depth description of the feature engineering techniques used to create the attributes.

Theory of Mind

In cognitive science, the Theory of mind (ToM) describes the ability to "understand the thoughts and feelings"[55] 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 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.

User Model

The User Model module (blue color in Figure 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 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 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 the 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.

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 listed in Section Feature Engineering, alongside with explanations on their numerical calculations.

An example representation of some dimensions that make up the model of the user can be seen in Figure 2. The Daily state submodule informs the Strategy Control (dark green color) module on emotions and what the focus topic of the daily mental health issues is (eg, are the mental issues connected more to the body or to thinking).

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 and decision-making. 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 5. The model holds the following psychological dimensions:

- Openness measures how inventive and curious people, as opposed to consistent and cautious, people are.
- Conscientiousness measures how efficient and organized people, as opposed to extravagant and careless, people are.
- Extraversion measures how outgoing and energetic people, as opposed to solitary and reserved, people are.
- Agreeableness measures how friendly and compassionate people, as opposed to critical and self-concerned, people are.
- Neuroticism measures how sensitive and nervous people, as opposed to resilient and confident, people are.

The system collects the numerical data necessary for the computational representation of the B5 modelling through a Finite State[81] conversational tree branch.

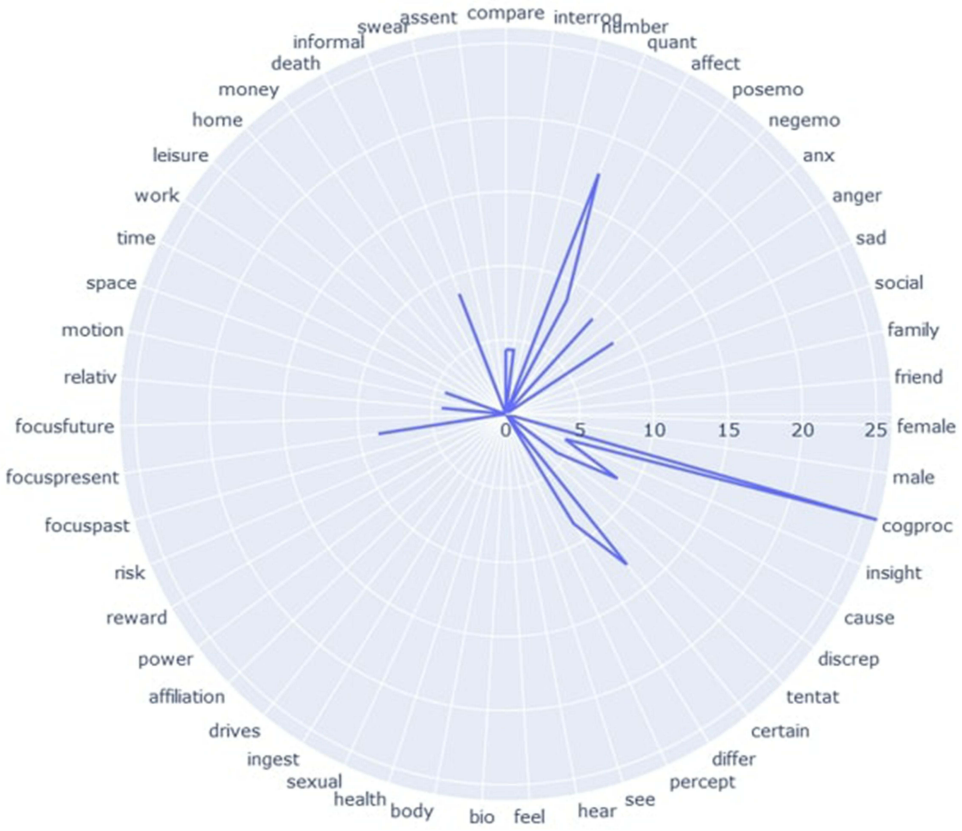

Figure 2 Example of a spider chart daily profile of a user. It contains several dimensions, based on the feature extraction.

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.[54] It is therefore ostensibly used to personalize the messages the system dispatches, eg, according to the dominant B5 dimension of the user,[83] by triggering the appropriate strategy from the Persuasion Strategies submodule in Expert & Domain Knowledge module (light green color).

SAD State

The SAD state submodule contains several ML models, including DT, BagDT, BoostDT, RF, CNB, kNN, MLP, LOG, and SVM (see Section Machine Learning Algorithms). They are trained to detect and forecast SAD levels and symptoms, described in Table 3, 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 Main Data Collection Study. The data can be collected by the system through dialogic questioning. For this, a longer time-series sample is required, which occurs if the conversation with a specific user does not occur just once.

The performance of the ML models can be seen in Section Computational Experiments.

Expert & Domain Knowledge

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.[84]

There are many techniques that are used in CBT, each with its own difficulty level (D1-D5), and *Supplementary Material: CBT Techniques* shows 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.

Persuasion Strategies

The submodule Persuasion Strategies contains the ontological knowledge on how to influence people with different psychological and personality characteristics. This makes, eg, 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.

The submodule's ontology comprises of a Domain Mapping Matrix (DMM)[85] between B5 dimensions and Cialdini's principles of persuasion (CPP).[86] 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, 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*.[87] 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 extracting a user's dimension with the highest value from the User Model and relating it to its CPP strategy counterpart in the ontology's Domain Mapping Matrix, to help with such strategy personalization as seen in the above example.

SAD Knowledge

The SAD knowledge submodule maps, using DMM, the topic of the mental health issue (eg, 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 found in *Supplementary Material: SAD-CBT-Mental Health Mapping* (see Table S17 for detailed mapping).

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.[88]

• The system uses different punctuations and emoticons depending on the user's mood. People in different mental states perceive sentence signs and symbols differently.[90]

Strategy Control

The Strategy control module (dark green in Figure 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 tracks the effectiveness of different strategies for a specific user (CBT techniques and persuasion strategies), if the same user continuously uses the system. It uses Ratio Formulas[90] 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.

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[90] 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.

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.

Natural Language Generation

The Natural language generation module (orange color in Figure 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[91] 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 Section Introduction. 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.

Large Language Model

The models below are the current natural language generation models used in this system. Because of their modular design, new models can be easily integrated. The selected model receives text based on the selected strategy and generates continuous text to enrich it. Thus, each output is unique. Numerous parameters allow for control over the enrichment of the original text. Through internal testing, the addition of two sentences appeared to generate the best outcomes.

• GPT-3: Generative Pre-trained Transformer 3 (GPT-3; stylized 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. The GPT-3 is capable of performing zero-shot, few-shot, and one-shot learning. Other OpenAI GPTs could also be used.

• 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 with between 125 million and 2.7 billion parameters on TensorFlow Research Cloud.

• GPT-J: GPT-J 6 B is a transformer model trained using Ben Wang's mesh transformer. "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 feedforward dimension of 16,384. The model dimensions were split into 16 heads, each with 256 dimensions. Rotary Position Embedding (RoPE) is applied to the 64 dimensions of each head. The model is trained with a tokenization vocabulary of 50257, using the same set of BPEs as GPTs.

• 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[92] with some modifications.[93]

GPT-J is the default selection in the system.

Natural Language Template

Contains Natural Language Templates[91] for conversational outputs.

Dialog Compilation

The submodule compiles the selected strategy from the Strategy control module with the enriched text from the Large language model relying on the Natural language template submodule.

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,[82] made with the Bad Bad Words dataset[94]

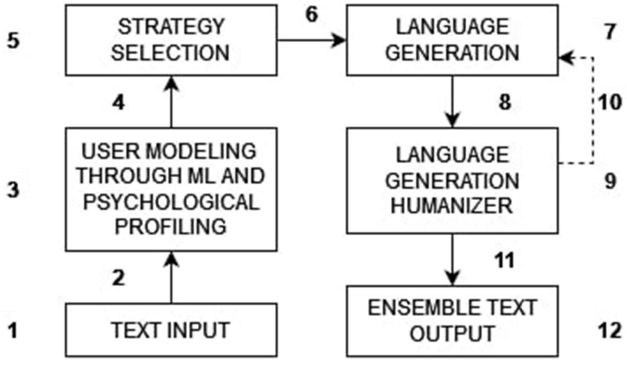

Figure 3 The system's pipeline through the modules in one conversational round.

and the Toxic Comments dataset,[95] as well as the Threshold-Filtering method in the ML detected negative sentiment scores using VADER (see Section Feature Engineering).

Operational Pipeline of the System

This Section presents the pipeline of the system executing one operational loop - one conversational round (see Figure 3). Subsequently, it provides an example of one such round. The steps in the conversation loop are the following:

- 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.
- The text is automatically processed to extract the features.
- 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.
- The user modelling profiling metrics are sent to the next part of the module.
- 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 text is sent to be augmented with the next part of the module.
- The text from the previous part is enriched by a generative pre-trained transformer based on a large language model (eg, GPT-3) with additional text. This makes responses more varied and alive for the user.
- The enriched text is passed to the Language generation humanizer.
- 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.
- Language generation humanizer returns the original text for another enrichment if deemed too risky for the user.
- The final text is compiled through a natural language template.
- Output of the final text.

A real-life example of the system in action can be seen in *Supplementary Material: System in Action Example*.

In summary, the cognitive architecture integrates several modules that work together to simulate theory of mind and deliver personalized mental health interventions. The system begins with Natural Language Processing, where user inputs are processed and analyzed for emotional content. This feeds into the Theory of Mind module, which includes a User Model that tracks daily mental states, personality traits, and SAD symptoms, along with forecasting capabilities. The Expert & Domain Knowledge module incorporates cognitive behavioral therapy (CBT) strategies and persuasion techniques, which are dynamically adjusted by the Strategy Control module to ensure tailored interventions. Finally, the Natural Language Generation module enhances the communication with users by generating personalized, context-aware responses. This architecture enables real-time detection and forecasting of mental health trends and provides adaptive, explainable interventions, setting the system apart from existing solutions.

Results

The Results section presents results from various experiments, focusing on computational experiments (see Section Computational Experiments for methodology used in them) of ML models for SAD state, and on the empirical interventional study (see Section Empirical Interventional Study for the research design) which compared our system with Woebot (see Section Proprietary Systems) for a description on Woebot as the currently best performing ICA for mental health support). The computational experiments demonstrated that this work's system outperformed SOTA models in detecting and forecasting SAD symptoms, achieving higher accuracy in most categories. The empirical study further validated these results, showing that this work's system provided more effective stress and anxiety relief compared to Woebot, though both systems were comparable in addressing depression. This system's forecasting ability, absent in other systems, was highlighted as its key strength.

Computational Experiments Results

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 Target Variables for descriptions on target variables). ML algorithms based on the Section Related Work were selected to be used in the fundamental experiments with SAD levels, and afterwards the best performing explainable ML algorithms were used subsequently. Section Machine Learning Algorithms provides methodological descriptions of the ML methods used.

Detection of SAD Levels and Symptoms from Single Text Entries

The whole dataset (see Section Dataset for more information on the dataset) was used to build the models. Features were extracted using the VADER and the LIWC frameworks (see Section Feature Engineering). The ML algorithms used for this task were RF, CNB, kNN, MLP, and LOG (see Section Machine Learning Algorithms). The 10-fold cross validation[96,97] with subject-wise splitting was used to evaluate the accuracy of the models. Majority class was used for the baseline model.

SAD Levels

Table 4 shows accuracies from various models detecting SAD levels from a single text entry.

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.

SAD Symptoms

Table 5 shows accuracies from kNN detecting SAD symptoms from a single text entry.

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

In this section, we present the computational performance results obtained from the models that utilize various ML algorithms. These models were designed to process individual text entries as inputs and forecast the levels of SAD (Table 6) and SAD (Table 7) symptoms that users will experience 7 days in the future. Features were extracted using the VADER and the LIWC frameworks (see Section Feature Engineering). The ML algorithms used for this task were RF, CNB, kNN, MLP, and LOG (see Section Machine Learning Algorithms). 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.

Quantitative Questionnaire Scores for 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 (see Table 8 for their accuracies) 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. The ML algorithms used for this task were kNN, LOG, CNB, CVM, DT, BagDT, BoostDT, and RF (see Section Machine Learning Algorithms). User-wise time series 10-fold cross

Table 4 Model accuracies in detecting SAD (stress, anxiety, and depression) levels from a single text entry.

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**

Notes: SAD (stress, anxiety, and depression) levels detection from a single text entry after the system's question on the user's daily mood, experiences, and events. The bold numbers represent the highest value in each SAD category for the given method. The methods used include: Random Forest (RF), Complement Naive Bayes (CNB), kNN (k-Nearest Neighbor), Multiple Layer Perceptron Classifier (MLP), and Logistic Regression (LOG).

Table 5 Accuracy of kNN in detecting individual SAD symptoms from a single text entry.

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

Notes: SAD (stress, anxiety, and depression) symptoms detection using kNN (k-Nearest Neighbor) for explainability from a single text entry after the system's question on the user's daily mood, experiences, and events.

Table 6 Accuracy of machine learning models in forecasting 7-day SAD (stress, anxiety, and depression) levels based on single text entries.

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

Notes: SAD (stress, anxiety, and depression) levels 7-day forecast from a single text entry after the system's question on the user's daily mood, experiences, and events. The bold numbers represent the highest value in each SAD category for the given method. The methods used include: Random Forest (RF), Complement Naive Bayes (CNB), kNN (k-Nearest Neighbor), Multiple Layer Perceptron Classifier (MLP), and Logistic Regression (LOG).

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 the 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.

The comparison between the models' performance from the above three computational experiments (detection of SAD levels and symptoms from single text entries, 7-day forecasting of SAD levels and symptoms from single text entries, and quantitative questionnaire scores for SAD level time series 7-day forecasting) and the systems reviewed in

Table 7 Accuracy of kNN in forecasting 7-day SAD (stress, anxiety, and depression) symptoms from single text entries.

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

Notes: SAD (stress, anxiety, and depression) symptoms 7-day forecast using kNN (k-Nearest Neighbor) for explainability from a single text entry after the system's question on the user's daily mood, experiences, and events.

Table 8 Accuracy of models in forecasting 7-day SAD (stress, anxiety, and depression) levels based on quantitative questionnaire scores.

SAD	Baseline	kNN	LOG	CNB	CVM	DT	BagDT	BoostDT	RF
Stress	53.30	77.23	77.23	77.23	**80.19**	67.33	76.24	74.26	79.21
Anxiety	73.70	88.12	**89.11**	87.13	**89.11**	79.21	89.11	84.16	86.14
Depression	66.60	**85.15**	82.18	84.16	84.16	84.16	83.17	82.18	84.16

Notes: 7-day forecast of SAD (stress, anxiety, and depression) level quantitative questionnaire scores. The bold numbers represent the highest value in each SAD category for the given method. The methods used include: Random Forest (RF), Complement Naive Bayes (CNB), kNN (k-Nearest Neighbor), Multiple Layer Perceptron Classifier (MLP), and Logistic Regression (LOG).

the section Related Work can be seen in Table 9 (see also section User Assessment Performance Comparison for a more detailed discussion on the comparison).

Empirical Interventional Study Results

This section presents the results of the empirical interventional study, described in Section Empirical Interventional Study. The study collected data from 42 participants and randomly sorted them into two groups. Group "Woebot" used Woebot (see Section Related Work) for a quick daily therapeutic check-in. Group "Our system" used the system

Table 9 Comparison of this work with prior studies in terms of assessed categories, reported accuracies, and best-performing machine learning methods.

System	Assessed Categories	Accuracies Reported	Best Methods
[38]	Depression Suicidal ideation Insomnia Hypersomnia Weight change Inappropriate guilt	detection: 0.91 (F1 score)	Random Forest: Logistic Regression
[39]	Sentiment emotions	detection: 81%	Fuzzy Matching
[40]	Valence arousal	detection: 65.7–82.4%	AdaBoost Regression
[44]	Seven emotions	detection: 84%	Long short-term memory (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%	k-Nearest Neighbor (kNN)
ChatGPT[97]	Stress Depression Suicidality	detection: 33–85%	Transformer

Notes: This table highlights the assessed categories, detection and forecast accuracies, and best-performing ML (machine learning) methods used in this work and prior studies. The comparison showcases the range of accuracies and methods applied across different systems.

described in this work (see Section Cognitive Architecture Design) for a quick daily therapeutic check-in. Stress, anxiety and depression were measured before and after the check-in.

Stress

Figure 4 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 ($t = 0.649$, $p = 0.642$). Sign testing determined that the stress score in the "Woebot" group did not change statistically significantly after the use ($p = 0.063$), while the stress score in the "Our system" group did change statistically significantly after the use ($p = 0.004$).

Table 10 provides a detailed summary of pre- and post-ICA usage stress scores for both groups, including statistical significance results.

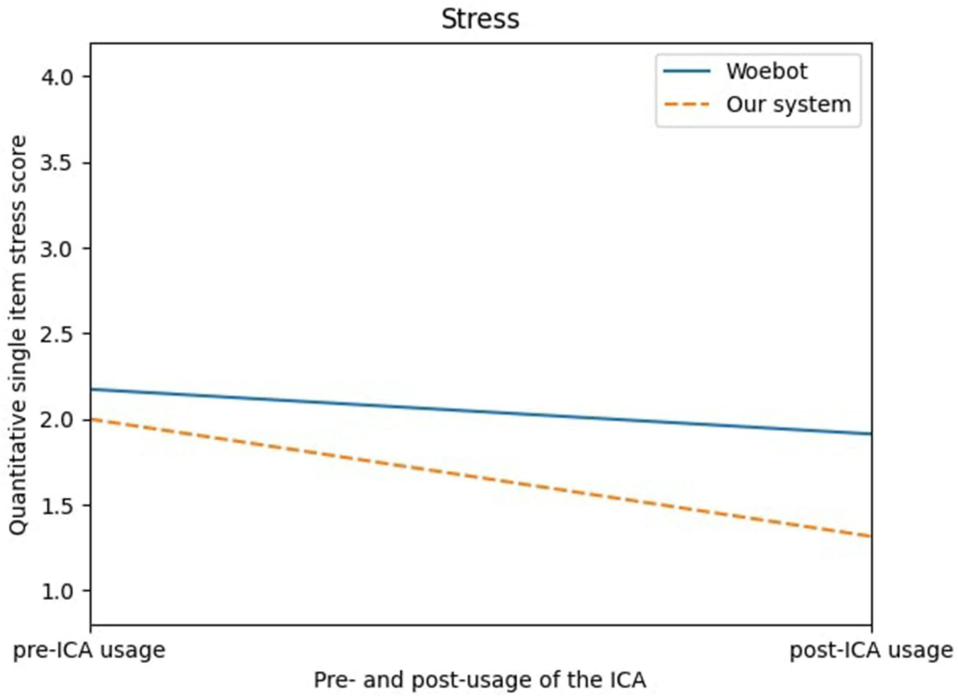

Figure 4 Plot comparing participant stress scores pre- and post-ICA usage in two different groups, "Woebot" and "Our system". Only participant stress in the "Our system" group changed statistically significantly.

Anxiety

Figure 5 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 ($t = 0.227$, $p = 0.822$). Sign testing determined that the anxiety score in the "Woebot" group did not change statistically significantly after the use ($p = 0.125$), while the anxiety score in the "Our system" group did change statistically significantly after the use ($p = 0.008$).

Table 11 provides a detailed summary of pre- and post-ICA usage anxiety scores for both groups, including statistical significance results.

Depression

Figure 6 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 ($t = 1.014$, $p = 0.317$). Sign testing determined that the depression score in neither the "Woebot" group ($p = 0.5$) nor in the "Our system" group ($p = 0.625$) changed statistically significantly after the use.

Table 10 Statistical comparison of stress scores, including pre-ICA usage group comparability and pre- vs. post-ICA usage changes, between the "Woebot" group and "Our system".

	Comparative Independent t-Testing for Pre-ICA Usage Group Comparability	**Pre- vs Post-ICA Usage Comparative Sign Testing**
Group "Woebot"	$t = 0.649$, $p = 0.642$	$p = 0.063$
Group "Our system"		$p = 0.004$

Notes: An independent samples *t*-test was conducted to assess the comparability of stress scores between the two groups before ICA usage in order to acquire t- (*t*) and p-values (*p*). No significant difference was found ($t = 0.649$, $p = 0.642$). Sign tests were performed to assess within-group changes in stress scores post-ICA usage, with a significant reduction observed in the "Our system" group ($p = 0.004$), but no significant change in the "Woebot" group ($p = 0.063$).

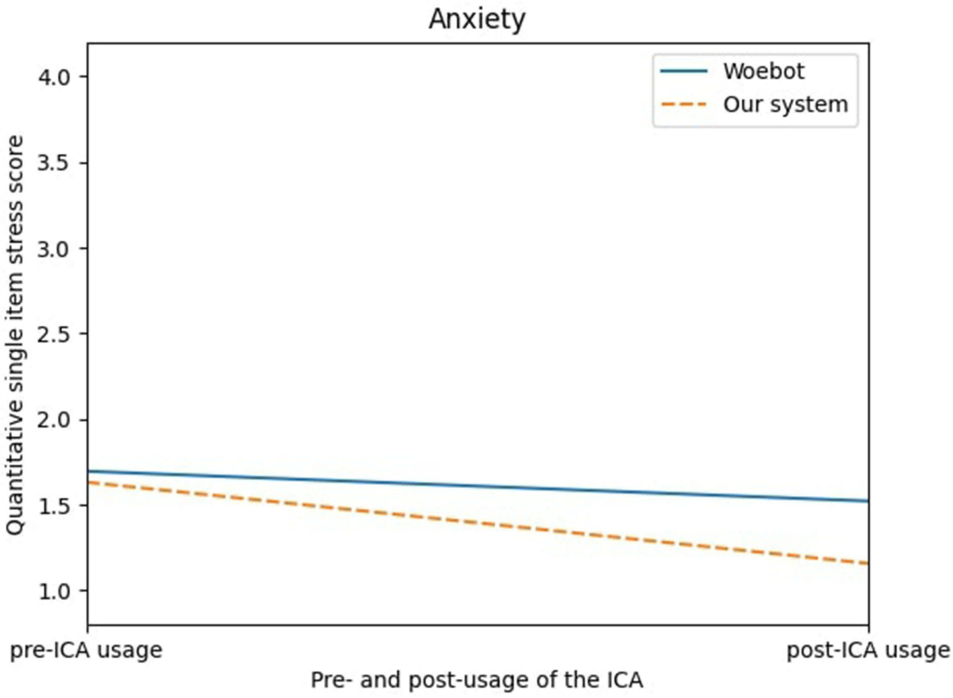

Figure 5 Plot comparing participant anxiety scores pre- and post-ICA usage in two different groups, "Woebot" and "Our system". Only participant anxiety in the "Our system" group changed statistically significantly.

Table 12 presents the pre- and post-ICA usage depression scores, showing the lack of statistically significant change in both groups.

User Experience Questionnaire Measures

In addition to 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:[80] *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 $M = 4.609$, $p = 0.084$).

Table 11 Statistical comparison of anxiety scores, including pre-ICA usage group comparability and pre- vs. post-ICA usage changes, between the "Woebot" group and "Our system".

	Comparative Independent t-Testing for Pre-ICA Usage Group Comparability	**Pre- vs Post-ICA Usage Comparative Sign Testing**
Group "Woebot"	$t = 0.227$, $p = 0.822$	$p = 0.125$
Group "Our system"		$p = 0.008$

Notes: An independent samples *t*-test was conducted to assess the comparability of anxiety scores between the two groups before ICA usage in order to acquire t- (*t*) and p-values (*p*). No significant difference was found ($t = 0.227$, $p = 0.822$). Sign tests were performed to assess within-group changes in anxiety scores post-ICA usage, with a significant reduction observed in the "Our system" group ($p = 0.008$), but no significant change in the "Woebot" group ($p = 0.125$).

Table 12 Statistical comparison of depression scores, including pre-ICA usage group comparability and pre- vs. post-ICA usage changes, between the "Woebot" group and "Our system".

	Comparative Independent t-Testing for Pre-ICA Usage Group Comparability	**Pre- vs Post-ICA Usage Comparative Sign Testing**
Group "Woebot"	t = 1.014, p = 0.317	p = 0.5
Group "Our system"		p = 0.625

Notes: An independent samples *t*-test was conducted to assess the comparability of depression scores between the two groups before ICA usage in order to acquire t- (t) and p-values (p). No significant difference was found (t = 1.014, p = 0.317). Sign tests assessed within-group changes in depression scores post-ICA usage, revealing no statistically significant changes in either the "Woebot" group (p = 0.5) or the "Our system" group (p = 0.625).

Discussion

This section compares this work's system to SOTA systems, described in Section Related Work. The evaluation is mostly represented using comparison tables. It furthermore discusses this work's hypothesis, this work's strengths and limitations, and finally offers some ideas of the authors for future work.

User Assessment Performance Comparison

The results of this work's system show significant improvements in detecting and forecasting SAD levels and symptoms compared to prior systems. Out of the nine extensively reviewed systems, four reported classification accuracies. These systems, along with the performance of this work's system, are presented in Table 9. The table shows that this work's system not only detects the most assessed categories but also achieves the highest accuracy among the compared systems. Specifically, this work's system achieved the highest accuracy of 91.41% for assessed SAD categories (see

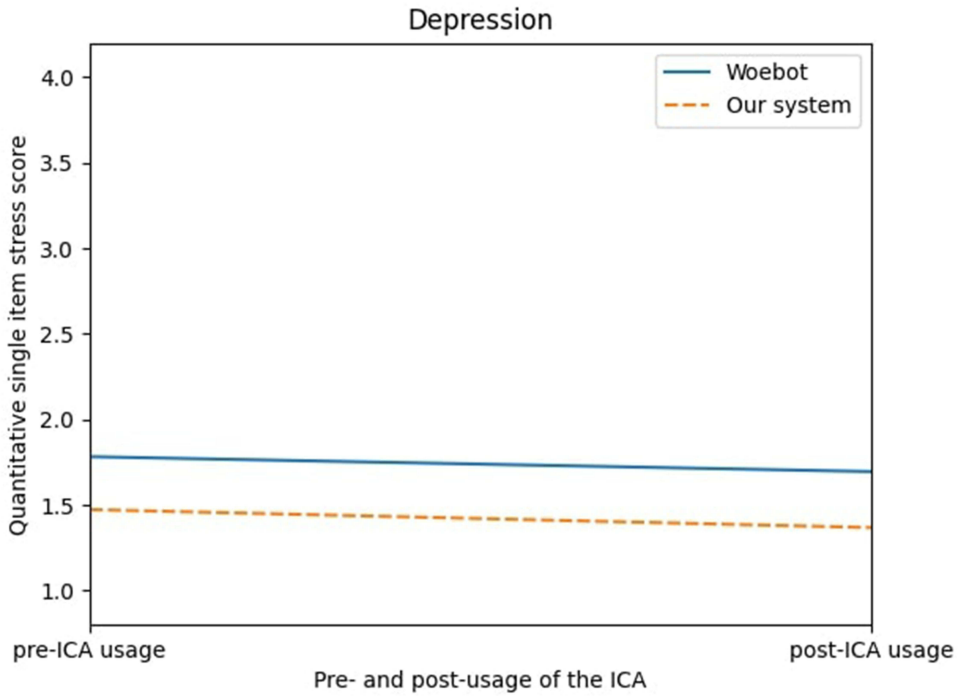

Figure 6 Plot comparing participant depression scores pre- and post-ICA usage in two different groups, "Woebot" and "Our system". Participant depression scores changed statistically significantly in neither of the groups.

Table 5, *Feeling hot or cold*), outperforming prior systems like Delahunty et al's sequence-to-sequence neural networks.[38] Furthermore, while systems like Denecke et al's SERMO[39] and Ghandeharioun et al's EMMA[40] employed lexicon-based emotion recognition and smartphone sensor data, respectively, they achieved lower classification performance compared to the ensemble of models used in this work's system.

A key differentiator of this work's system is its ability to forecast mental health trends up to 7 days in advance (as shown in Tables 6–8), a feature absent in all reviewed state-of-the-art (SOTA) systems. None of the other systems, including Woebot[46] and Tess,[45] have forecasting capabilities; they focus solely on detecting the current mental state of the user based on real-time input. For example, while Woebot has been effective in reducing anxiety and depression symptoms in the short term, it lacks the proactive intervention potential that comes with forecasting. This work's system, by contrast, achieved the forecasting accuracy range of 71.54–87.68% for assessed SAD categories (see Tables 6–8), allowing it to provide anticipatory support to users (discussed further in the following Section).

Finally, ChatGPT has recently been evaluated on its classification capabilities for stress, depression, and suicidality, with reported accuracies of 73%, 85%, and 33%, respectively.[97] This work's system outperformed ChatGPT in two out of three categories, with accuracies of 74.56% for stress, 83.33% for depression, and 76.20% for suicidality (as seen in Tables 4–5). This demonstrates the robustness of this work's system in handling domain-specific mental health tasks, a key improvement over general-purpose conversational agents like ChatGPT, which tend to struggle in specialized contexts.

Evaluation of System's Effectiveness

The effectiveness of this system was evaluated through an empirical interventional study (see Section Empirical Interventional Study Results), in which it was directly compared to Woebot, the current SOTA system for mental health support. Woebot is frequently cited in research for its interventional success in reducing stress, anxiety, and depression symptoms, and its effectiveness has been replicated across various studies. However, as shown in our empirical study, this work's system performed better than Woebot in reducing both stress and anxiety in participants (see Figure 4–5). Specifically, while the Woebot group did not achieve statistically significant changes in stress or anxiety, this work's system produced statistically significant reductions ($p = 0.048$ for stress; $p = 0.040$ for anxiety).

Neither system succeeded in statistically significantly reducing depression in participants, as reported in Section Empirical Interventional Study Results (see Figure 6). This aligns with previous findings that depression is often more difficult to alleviate in short-term interventions. However, this work's system offers a distinct advantage with its forecasting capability, allowing it to predict the likelihood of future depressive episodes and potentially intervene before symptoms worsen, which is a feature Woebot lacks. While Woebot has demonstrated effectiveness in short-term symptom relief, it does not possess the advanced cognitive architecture required to forecast trends in mental health, thus limiting its proactive support.

In addition to clinical outcomes, this work's system was evaluated more favorably than Woebot on supportiveness (as measured by the User Experience Questionnaire), with participants rating this work's system as significantly more supportive ($M = 5.368$) than Woebot ($M = 4.261$, $p = 0.041$). This result highlights the importance of personalized and adaptive responses that leverage the system's theory of mind simulation to better understand and respond to individual users. Both systems scored similarly on the usual-leading edge metric, reflecting comparable levels of technological novelty and advancement.

The results from this study suggest that the integration of theory of mind and personalized strategies in this work's system provides an effective solution for stress and anxiety management.

Qualitative Evaluation from the Empirical Interventional Study

In the empirical interventional study, 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 number 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. These findings suggest that future ICA development should aim to combine our system's robust cognitive

architecture with the more polished user interfaces found in proprietary systems like Woebot to enhance user engagement and satisfaction.

Broader Comparison with Related Systems

In comparison to previous systems, such as Delahunty et al and Denecke et al, our approach represents a more holistic solution by integrating cognitive architecture and behavior change models, which are further enhanced by machine learning-driven personalization. While previous works focused on detecting symptoms or providing basic interventions, none combined these functionalities with long-term forecasting and adaptive response generation. For example, systems like EMMA and Bonobot[43] lack the real-time adaptation and prediction capabilities that our system provides, limiting their effectiveness in addressing dynamic and fluctuating mental health states over time.

Hypothesis Testing

Hypothesis H in this work proposed that

> an intelligent cognitive assistant for attitude and behavior change for stress, anxiety, and depression will achieve comparable state-of-The-art results if it simulates theory of mind in a novel artificial cognitive architecture.

It may be concluded that the hypothesis H was confirmed based on the User assessment performance comparison (Section User assessment performance comparison), Evaluation of the system's effectiveness (Section Evaluation of the system's effectiveness), and Qualitative evaluation from the empirical interventional study (Section Qualitative evaluation from the empirical interventional study).

Strengths and Limitations

This work introduces several key strengths that advance the field of intelligent cognitive assistants for mental health support. A primary strength is the novel cognitive architecture simulating theory of mind, allowing the system to model and understand user mental states with high precision. This enables personalized interventions that adapt to each user's emotional and cognitive profile, offering deeper personalization compared to existing systems, which rely on rigid models.

Another notable strength is the system's forecasting capability, predicting stress, anxiety, and depression symptoms 7 days in advance. This proactive approach allows for anticipatory interventions, surpassing the immediate detection focus of previous systems, and offering a more comprehensive solution to mental health challenges.

The system's use of explainable machine learning models, such as k-nearest neighbors, is another advantage. By providing interpretable outputs, it ensures that mental health professionals can understand and trust its recommendations, unlike black-box models used in some other ICAs.

Finally, the development of a novel dataset combining quantitative and qualitative data from daily diaries and diagnostic-level questionnaires is a significant contribution. This dataset, with over 1000 instances, supports robust machine learning models and addresses the scarcity of high-quality, domain-specific data in mental health research.

In assessing the findings of this work, it is also crucial to recognize several notable limitations. First, 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 are excluded from the sample. Second, 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. Third, the target attributes are subject to change as definitions of mental health phenomena evolve over time, which may consequently impact the models' detection and forecasting performance. Fourth, 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.

Future Work

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 (eg, LSTM); as well as potential aforementioned expansion of the collected dataset. The second encompasses a non-quasi-experimental empirical interventional study with a higher number 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 the participants.

Conclusion

This work presents a comprehensive computational psychotherapy system for mental health prediction and behavior change using a conversational agent. The system makes two key contributions. First, it introduces a novel, golden standard dataset, including panel data with quantitative SAD symptom scores and qualitative daily diary entries, addressing a gap in computational resources for psychotherapeutic and psychiatric research. Second, it offers a system for SAD symptom relief, primarily driven by a simulated Theory of Mind (ToM), which models user cognition through a combination of psychological and cognitive user modeling, mental health ontologies, machine learning detection and forecasting models, and behavior change prompt generators.

The hypothesis that an intelligent cognitive assistant with ToM would achieve results comparable to or better than state-of-the-art systems was confirmed. The system outperformed other systems in detecting a wider range of mental health phenomena, achieved higher detection accuracies, and uniquely forecasted mental health trends, which others could not. Additionally, it was more effective in supporting participants compared to Woebot; however, Woebot offered a more polished interface, while this system remains a research prototype.

These contributions provide both a robust dataset and a novel computational psychotherapeutic system for symptom relief, establishing a benchmark for intelligent, adaptive interventions with the potential to improve predictive accuracy and therapeutic outcomes in future mental health applications.

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.

Funding

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P2-0209 and Young researchers postgraduate research funding).

Disclosure

Dr Tine Kolenik reports grants from Slovenian Research Agency during the conduct of the study. The authors report no conflicts of interest in this work.

References

1. Ornell F, Borelli WV, Benzano D, et al. The next pandemic: impact of COVID-19 in mental healthcare assistance in a nationwide epidemiological study. *Lancet Reg Health – Am*. 2021;4:100061. doi:10.1016/j.lana.2021.100061
2. United Nations Sustainable Development – 17 Goals to Transform Our World. 2020. https://www.un.org/sustainabledevelopment/.
3. Winkler P, Krupchanka D, Roberts T, 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*. 2017;4(8):634–642. doi:10.1016/S2215-0366(17)30135-9
4. Mental Health Foundation. *Stress: Are We Coping?* Mental Health Foundation; 2018.
5. Baxter AJ, Scott KM, Vos T, Whiteford HA. Global prevalence of anxiety disorders: a systematic review and meta-regression. *Psychol Med*. 2013;43(5):897–910. doi:10.1017/S003329171200147X
6. Twenge J. Time Period and Birth Cohort Differences in Depressive Symptoms in the U.S. 1982–2013. *Soc Indic Res Int Interdiscip J Qual–Life Meas*. 2015;121(2):437–454. doi:10.1007/s11205-014-0647-1
7. Wang PS, Aguilar-Gaxiola S, Alonso J, et al. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *Lancet*. 2007;370(9590):841–850. doi:10.1016/S0140-6736(07)61414-7.
8. Schmidtke A, Bille-Brahe U, DeLeo D, 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*. 1996;93(5):327–338. doi:10.1111/j.1600-0447.1996.tb10656.x.
9. Curtin SC, Warner M, Hedegaard H. Increase in Suicide in the United States, 1999-2014. *NCHS Data Brief*. 2016;2016(241):1–8.
10. Auerbach J, Miller BF. COVID-19 Exposes the Cracks in Our Already Fragile Mental Health System. *Am J Public Health*. 2020;110(7):969–970. doi:10.2105/AJPH.2020.305699
11. World Health Organization. *Investing in Mental Health*. World Health Organization; 2003.
12. Kolenik T, Gams M. Persuasive Technology for Mental Health: one Step Closer to (Mental Health Care) Equality? *IEEE Technol Soc Mag*. 2021;40(1):80–86. doi:10.1109/MTS.2021.3056288
13. Mindu T, Mutero IT, Ngcobo WB, Musesengwa R, Chimbari MJ. Digital Mental Health Interventions for Young People in Rural South Africa: prospects and Challenges for Implementation. *Int J Environ Res Public Health*. 2023;20(2):1453. doi:10.3390/ijerph20021453
14. Moutoussis M, Shahar N, Hauser TU, Dolan RJ. Computation in Psychotherapy, or How Computational Psychiatry Can Aid Learning-Based Psychological Therapies. *Comput Psychiatr*. 2018;2:50–73. doi:10.1162/CPSY_a_00014
15. Gams M, Kolenik T. Relations between Electronics, Artificial Intelligence and Information Society through Information Society Rules. *Electronics*. 2021;10(4):514. doi:10.3390/electronics10040514
16. Fogg BJ. *Persuasive Technology: Using Computers to Change What We Think and Do*. Morgan Kaufmann Publishers Inc.; 2003.
17. Orji R, Moffatt K. Persuasive technology for health and wellness: state-of-The-art and emerging trends. *Health Informatics J*. 2018;24(1):66–91. doi:10.1177/1460458216650979.
18. Oakley J Intelligent Cognitive Assistants (ICA). 2018. Accessed Nov 28, 2025. Available from: https://www.nsf.gov/crssprgm/nano/reports/ICA2_Workshop_Report_2018.pdf.
19. Cognitive Architecture 2020. Accessed Nov 28, 2024. Available from: http://cogarch.ict.usc.edu/.
20. Garrod S, Pickering MJ. Why is conversation so easy? *Trends Cognit Sci*. 2004;8(1):8–11. doi:10.1016/j.tics.2003.10.016.
21. Bakker D, Kazantzis N, Rickwood D, Rickard N. Mental Health Smartphone Apps: review and Evidence-Based Recommendations for Future Developments. *JMIR Ment Health*. 2016;3(1):e7. doi:10.2196/mental.4984
22. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inf Assoc*. 2018;25(9):1248–1258. doi:10.1093/jamia/ocy072
23. Montenegro JLZ, da CCA, da R RR. Survey of conversational agents in health. *Expert Syst Appl*. 2019;129:56–67. doi:10.1016/j.eswa.2019.03.054.
24. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied Conversational Agents in Clinical Psychology: a Scoping Review. *J Med Internet Res*. 2017;19(5):e151. doi:10.2196/jmir.6553
25. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and Conversational Agents in Mental Health: a Review of the Psychiatric Landscape. *Can J Psychiatry*. 2019;64(7):456–464. doi:10.1177/0706743719828977
26. Kolenik T. Methods in Digital Mental Health: smartphone-based Assessment and Intervention for Stress, Anxiety and Depression. In: Comito C, Forestiero A, Zumpano E, editors. *Integrating Artificial Intelligence and IoT for Advanced Health Informatics*. Springer; 2021.
27. Thorp HH. ChatGPT is fun, but not an author. *Science*. 2023;379(6630):313. doi:10.1126/science.adg7879
28. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *Npj Digit Med*. 2021;4(1):93. doi:10.1038/s41746-021-00464-x
29. Pritchard DJ, Hurly TA, Tello-Ramos MC, Healy SD. Why study cognition in the wild (and how to test it)? *J Exp Anal Behav*. 2016;105(1):41–55. doi:10.1002/jeab.195
30. Prince M. 9 - Epidemiology. In: Wright P, Stern J, Phelan M, editors. *Core Psychiatry*. Third Edition ed. W.B. Saunders; 2012:115–129. doi:10.1016/B978-0-7020-3397-1.00009-4.
31. van Berkel N *Data Quality and Quantity in Mobile Experience Sampling*. PhD Thesis. 2019. http://hdl.handle.net/11343/227682.
32. Kubiak T, Smyth JM. Connecting Domains—Ecological Momentary Assessment in a Mobile Sensing Framework. In: Baumeister H, Montag C editors. *Digital Phenotyping and Mobile Sensing: New Developments in Psychoinformatics*. Springer International Publishing; 2019:201–207. doi:10.1007/978-3-030-31620-4_12.
33. Colombo D, Fernández-álvarez J, Patané A, et al. Current State and Future Directions of Technology-Based Ecological Momentary Assessment and Intervention for Major Depressive Disorder: a Systematic Review. *J Clin Med*. 2019;8(4):465. doi:10.3390/jcm8040465

34. Kansoun Z, Boyer L, Hodgkinson M, Villes V, Lançon C, Fond G. Burnout in French physicians: a systematic review and meta-analysis. *J Affect Disord*. 2019;246:132–147. doi:10.1016/j.jad.2018.12.056
35. Kolenik T, Gams M. Intelligent Cognitive Assistants for Attitude and Behavior Change Support in Mental Health: state-of-The-Art Technical Review. *Electronics*. 2021;10(11):1250. doi:10.3390/electronics10111250
36. Kolenik T, Gams M. *PerMEASS – Personal Mental Health Virtual Assistant with Novel Ambient Intelligence Integration*. 2020:8–12.
37. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol*. 2005;8(1):19–32. doi:10.1080/1364557032000119616
38. Petticrew M, Roberts H. *Systematic Reviews in the Social Sciences: A Practical Guide*. Wiley; 2008. https://books.google.si/books?id=ZwZ1_xU3E80C.
39. Delahunty F, Wood ID, Arcan M 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*.; 2018:327–338.
40. Denecke K, Vaaheesan S, Arulnathan A A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test. *IEEE Trans Emerg Top Comput*. 2020:1. doi:10.1109/TETC.2020.2974478.
41. Ghandeharioun A, McDuff D, Czerwinski M, Rowan K EMMA: an Emotion-Aware Wellbeing Chatbot. In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*; 2019:1–7. doi:10.1109/ACII.2019.8925455.
42. Khadikar S, Sharma P, Paygude P. Compassion Driven Conversational Chatbot Aimed for better Mental Health. *Zeich J*. 2020;6(9):121–127.
43. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an Artificially Empathic Conversational Agent for Mental Health Applications: system Design and User Perceptions. *J Med Internet Res*. 2018;20(6):e10148. doi:10.2196/10148.
44. Park S, Choi J, Lee S, et al. Designing a Chatbot for a Brief Motivational Interview on Stress Management: qualitative Case Study. *J Med Internet Res*. 2019;21(4):e12231. doi:10.2196/12231
45. Pola S, Chetty MSR. Behavioral therapy using conversational chatbot for depression treatment using advanced RNN and pretrained word embeddings. *Mater Today Proc*. 2021. doi:10.1016/j.matpr.2021.02.521.
46. Rishabh C, Anuradha J. COUNSELLOR CHATBOT. *Int Res J Comput Sci*. 2018;3(5):126–136.
47. contributors W. ELIZA — Wikipedia, The Free Encyclopedia. 2021. https://en.wikipedia.org/w/index.php?title=ELIZA&oldid=1012889844.
48. Wikipedia contributors. Artificial Linguistic Internet Computer Entity — Wikipedia, The Free Encyclopedia. 2020. https://en.wikipedia.org/w/index.php?title=Artificial_Linguistic_Internet_Computer_Entity&oldid=993396811.
49. Yorita A, Egerton S, Oakman J, Chan C, Kubota N A Robot Assisted Stress Management Framework: using Conversation to Measure Occupational Stress. In: *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2018:3761–3767.
50. Yorita A, Egerton S, Chan C, Kubota N Chatbot for Peer Support Realization based on Mutual Care. In: *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*.; 2020:1601–1606. doi:10.1109/SSCI47803.2020.9308277.
51. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. JMIR Ment Health. 2018;5(4):e64. doi:10.2196/mental.9782
52. Fitzpatrick KK, Darcy A, Vierhile M. 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*. 2017;4(2):e19. doi:10.2196/mental.7785
53. Prochaska JJ, Vogel EA, Chieng A, et al. A Therapeutic Relational Agent for Reducing Problematic Substance Use (Woebot): development and Usability Study. *J Med Internet Res*. 2021;23(3):e24850. doi:10.2196/24850
54. Schiepek G, Gelo O, Viol K, et al. Complex individual pathways or standard tracks? A data-based discussion on the trajectories of change in psychotherapy. *Couns Psychother Res*. 2020;20(4):689–702. doi:10.1002/capr.12300
55. Hirsh JB, Kang SK, Bodenhausen GV. Personalized Persuasion: tailoring Persuasive Appeals to Recipients' Personality Traits. *Psychol Sci*. 2012;23(6):578–581. doi:10.1177/0956797611436349
56. Leslie AM, Friedman O, German TP. Core mechanisms in 'theory of mind. *Trends Cognit Sci*. 2004;8(12):528–533. doi:10.1016/j.tics.2004.10.001
57. Brown TB, Mann B, Ryder N, et al. Language Models are Few-Shot Learners. *CoRR*. 2020;14165.
58. Darcy A, Beaudette A, Chiauzzi E, et al. Anatomy of a Woebot® (WB001): agent guided CBT for women with postpartum depression. *Expert Rev Med Devices*. 2022;19(4):287–301. doi:10.1080/17434440.2022.2075726
59. Demirci HM *User Experience over Time with Conversational Agents: case Study of Woebot on Supporting Subjective Well-Being*. Master's Thesis. Middle East Technical University; 2018.
60. Jain A, Patel H, Nagalapatti L, 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*. KDD '20. Association for Computing Machinery; 2020:3561–3562. doi:10.1145/3394486.3406477.
61. Gupta N, Mujumdar S, Patel H, et al. Data Quality for Machine Learning Tasks. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. KDD '21. Association for Computing Machinery; 2021:4040–4041. doi:10.1145/3447548.3470817.
62. Schiepek G, Eckert H, Aas B, Wallot S, Wallot A. *Integrative Psychotherapy a Feedback-Driven Dynamic Systems Approach*. Hogrefe Verlag GmbH & Co. KG; 2015; doi:10.1027/00472-000.
63. Rammstedt B, John OP. Measuring personality in one minute or less: a 10-item short version of the Big Five Inventory in English and German. *J Res Personal*. 2007;41(1):203–212. doi:10.1016/j.jrp.2006.02.001
64. Olden M, Holle R, Heid IM, Stark K. IDGenerator: unique identifier generator for epidemiologic or clinical studies. *BMC Med Res Methodol*. 2016;16(1):120. doi:10.1186/s12874-016-0222-3.
65. Lovibond SH, Lovibond PF. *Manual for the Depression Anxiety Stress Scales*. Psychology Foundation of Australia; 1996. https://books.google.si/books?id=mXoQHAAACAAJ.
66. Beck AT, Epstein N, Brown G, Steer R. Beck anxiety inventory. *J Consult Clin Psychol*. 1993;1993:1.
67. Beck AT, Steer RA, Brown GK, et al. *Beck Depression Inventory*. Harcourt Brace Jovanovich New York:; 1987.
68. Ratcliffe M. *Experiences of Depression: A Study in Phenomenology*. Oxford University Press; 2015. https://books.google.si/books?id=0UePBQAAQBAJ.
69. Orne MT. Demand characteristics. In: *Introducing Psychological Research*. Springer; 1996:395–401.
70. Friedman D Machine Learning from Scratch. *GitHub Repos*. 2020. https://dafriedman97.github.io/mlbook/content/introduction.html.
71. Seref B, Bostanci E Performance comparison of Naïve Bayes and complement Naïve Bayes algorithms. In: *2019 6th International Conference on Electrical and Electronics Engineering (ICEEE)*. IEEE; 2019:131–138.

72. Peterson LE. K-nearest neighbor. *Scholarpedia*. 2009;4(2):1883. doi:10.4249/scholarpedia.1883.
73. Werbos PJ. Backpropagation through time: what it does and how to do it. *Proc IEEE*. 1990;78(10):1550–1560. doi:10.1109/5.58337
74. Noble WS. What is a support vector machine? *Nat Biotechnol*. 2006;24(12):1565–1567. doi:10.1038/nbt1206-1565.
75. Hutto C, Gilbert E 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*. 2014;8:216–225.
76. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. *J Lang Soc Psychol*. 2010;29 (1):24–54. doi:10.1177/0261927X09351676.
77. Spitzer RL, Kroenke K, Williams JB, Löwe B. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Arch Intern Med*. 2006;166 (10):1092–1097. doi:10.1001/archinte.166.10.1092.
78. DiNardo J. Natural Experiments and Quasi-Natural Experiments. In: *The New Palgrave Dictionary of Economics*. UK: Palgrave Macmillan;2016:1–12. doi:10.1057/978-1-349-95121-5_2006-1.
79. Arapovic-Johansson B, Wåhlin C, Kwak L, Björklund C, Jensen I. Work-related stress assessed by a text message single-item stress question. *Occup Med*. 2017;67(8):601–608. doi:10.1093/occmed/kqx111
80. Schrepp M, Hinderks A, Thomaschewski J. *Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios*. 2014:383–392. doi:10.1007/978-3-319-07668-3_37
81. Jokinen K, McTear M. *Spoken Dialogue Systems*. Springer International Publishing; 2010.
82. Grosan C, Abraham A. *Intelligent Systems*. Springer; 2011.
83. Jankovič A, Kolenik T, Pejović V. Can Personalization Persuade? Study of Notification Adaptation in Mobile Behavior Change Intervention Application. *Behav Sci*. 2022;12(5):116. doi:10.3390/bs12050116
84. Wikipedia. Cognitive behavioral therapy — Wikipedia, The Free Encyclopedia. 2022. https://en.wikipedia.org/wiki/Cognitive_behavioral_therapy.
85. Schmidt PP, Fay A. Applying the Domain-Mapping-Matrix to Identify the Appropriate Level of Detail of Simulation Models for Virtual Commissioning. *IFAC-Pap*. 2015;48(10):69–74. doi:10.1016/j.ifacol.2015.08.110
86. Cialdini RB *Influence: science and Practice*. Pearson Education; 2009. Available from: http://www.amazon.co.uk/Influence-Practice-Robert-B-Cialdini/dp/0205663788.
87. Alslaity A, Tran T The Effect of Personality Traits on Persuading Recommender System Users. In: *IntRS'20-Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*.; 2020:48–56.
88. Khanahmadi M, Malmir M, Eskandari H, Orang T. Evaluation of Visual Information Processing Speed in Depressed People. *Iran J Public Health*. 2013;42(11):1266–1273.
89. Hovermale O *1 Individual Differences in the Perception of Emoji: effects of Depression and Self-Esteem*. Master's Thesis. Ball State University Muncie, Indiana, USA; 2020.
90. Wikipedia contributors. Ratio — Wikipedia, The Free Encyclopedia. 2022. https://en.wikipedia.org/w/index.php?title=Ratio&oldid=1127160304.
91. van Deemter K, Krahmer E, Theune M. Squibs and Discussions: real versus Template-Based Natural Language Generation: a False Opposition? *Comput Linguist*. 2005;31(1):15–24. doi:10.1162/0891201053630291
92. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need; 2023.
93. Radford A, Narasimhan K. *Improving Language Understanding by Generative Pre-Training*. 2018.
94. Bad Bad Words dataset. 2020. https://www.kaggle.com/datasets/nicapotato/bad-bad-words.
95. Toxic Comments dataset. 2020. https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data.
96. Fushiki T. Estimation of prediction error by using K-fold cross-validation. *Stat Comput*. 2011;21(2):137–146. doi:10.1007/s11222-009-9153-8
97. Lamichhane B. *Evaluation of ChatGPT for NLP-Based Mental Health Applications*. 2023.