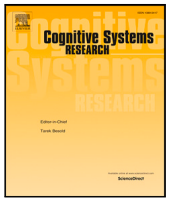
Cognitive Systems Research 85 (2024) 101203

Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com/locate/cogsys



A formal understanding of computational empathy in interactive agents

Andreas Brännström [a,*], Joel Wester [b], Juan Carlos Nieves [a]

[a] Umeå University, Department of Computing Science, Umeå, SE-90187, Sweden
[b] Aalborg University, Department of Computer Science, Aalborg, DE-9220, Denmark

ARTICLE INFO

Action editor: Alessandra Sciutti

Keywords:
Computational empathy
Conversational agents
Human–agent interaction
Knowledge engineering

ABSTRACT

Interactive software agents, such as chatbots, are progressively being used in the area of health and well-being. In such applications, where agents engage with users in interpersonal conversations for, e.g., coaching, comfort or behavior-change interventions, there is an increased need for understanding agents' empathic capabilities. In the current state-of-the-art, there are no tools to do that. In order to understand empathic capabilities in interactive software agents, we need a precise notion of empathy. The literature discusses a variety of definitions of empathy, but there is no consensus of a formal definition. Based on a systematic literature review and a qualitative analysis of recent approaches to empathy in interactive agents for health and well-being, a formal definition—an ontology—of empathy is developed. We present the potential of the formal definition in a controlled user-study by applying it as a tool for assessing empathy in two state-of-the-art health and well-being chatbots; Replika and Wysa. Our findings suggest that our definition captures necessary conditions for assessing empathy in interactive agents, and how it can uncover and explain trends in changing perceptions of empathy over time. The definition, implemented in Web Ontology Language (OWL), may serve as an automated tool, enabling systems to recognize empathy in interactions—be it an interactive agent evaluating its own empathic performance or an intelligent system assessing the empathic capability of its interlocutors.

1. Introduction

In everyday interpersonal interaction, humans largely depend on various notions in order to communicate. A cornerstone is the notion of empathy, the ability to understand and share the feelings of another (Elliott, Bohart, Watson, & Greenberg, 2011). Humans rely on their, as well as others', empathy to develop social bonds, to build and maintain trust, to collaborate, and to reach individual and shared goals. Humans need, and expect, empathy from others to different degrees in various contexts, but in scenarios where an individual seeks understanding or help from others, empathy is of particular concern.

Interactive software agents, embodied as Conversational Agents (CAs), such as chatbots (Brandtzaeg & Følstad, 2018), are progressively being used in the area of health and well-being (Luo, Aguilera, Lyles, & Figueroa, 2021), designed for, e.g., coaching, comforting and behavior-change applications (Schulman, Bickmore, & Sidner, 2011). Such systems, more commonly designed using various techniques from Natural Language Processing (NLP) (Beredo, Bautista, Cordel, & Ong, 2021; Duan, Zhao, Zhou, Qiu, & Liu, 2020) and Machine Learning (ML) (Li, Galley, Brockett, Spithourakis, Gao, & Dolan, 2016; Shang, Lu, & Li, 2015; Shumanov & Johnson, 2021) to build response generation models, are able to process and respond to user inputs in a conversational manner. Moreover, over the past few years, significant advancements have been made in the area of Large Language Models (LLMs) like GPT (Team OpenAI, 2022), T5 (Ni, Ábrego, Constant, Ma, Hall, Cer, & Yang, 2021), and LaMDA (Thoppilan, De Freitas, Hall, Shazeer, Kulshreshtha, Cheng, Jin, Bos, Baker, Du, et al., 2022), showcasing chatbots with remarkable ability to generate human-like language, yet with inherent limitations in sustaining focus and coherence during prolonged interactions (Ray, 2023). Nevertheless, these innovations have ignited interest in diverse chatbot applications. CAs, designed to simulate the nature of human–human conversations, have great potential in various contexts, for example, as previously highlighted, as social companions to promote mental well-being, helping individuals with loneliness, depression, and stress management. In order to provide such support, there is a prominent demand for systems to meet the users' need for relevance, coherent interactions and personalized support. This requires interactive agents able to perceive, understand and act empathically. When applications, in addition, are designed to provide intimate in-person interactions, such as in the setting of social companions, there is an increased need for the system to provide relatedness and interpersonal connectivity with its users, requiring a higher level of empathic capability. Given the strengths and limitations of current CAs, we seek to understand empathy in these systems.

* Corresponding author.
E-mail addresses: andreas.brannstrom@umu.se (A. Brännström), joelw@cs.aau.dk (J. Wester), juan.carlos.nieves@umu.se (J.C. Nieves).

https://doi.org/10.1016/j.cogsys.2023.101203
Received 13 May 2022; Received in revised form 30 September 2023; Accepted 14 December 2023
Available online 29 December 2023

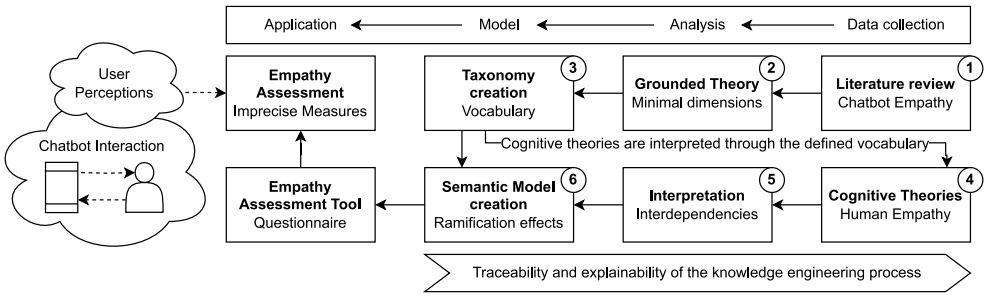


Fig. 1. Research design and methodology.

A starting point when aiming to develop empathic software agents is to learn more about how current interactive agents handle the empathic aspects of interaction, by assessing how the agents' empathic capabilities are perceived in interaction. A second challenge is to make these empathy assessments computational, allowing intelligent systems to assess and reason about empathic performance. In the research area of computational empathy (Boukricha, Wachsmuth, Carminati, & Knoeferle, 2013; Lowmanstone, 2021; Paiva, Leite, Boukricha, & Wachsmuth, 2017; Yalçın, 2019), research has largely focused on emotion recognition and empathic emotion expression, e.g., models for enabling agents to express empathic emotions based on facial mimicry (Boukricha et al., 2013). While emotion is a key component of empathy, it is not sufficient for understanding empathic capability from a holistic perspective. Different approaches for evaluating and assessing empathic capability in interactive agents have been proposed (Paiva et al., 2017; Yalçın, 2019). These approaches suggest metrics which typically are on an abstract level, not capturing explicit necessary conditions for perceived empathy. Furthermore, these metrics are not aimed at being computational and to be applied for automated empathy reasoning. In order to understand computational empathy from a holistic perspective in the state-of-the-art software systems and to approach automated empathic reasoning for the next generation of empathic software systems, we need computational models and tools for explaining and assessing perceptions of a system's empathic capabilities. For that, we need a formal definition and vocabulary, capturing the relevant concepts of empathy from a human–agent interaction perspective. The literature discusses a variety of (in)formal notions of human empathy (Decety & Jackson, 2004; Elliott, Bohart, Watson, & Murphy, 2018; Goldman, 2011; Preston, 2007; Szanto & Krueger, 2019; Watt, 2005), but there is no consensus on a formal definition of "computational empathy" for understanding interactive agents' empathic capabilities.

By considering the lacked consensus on empathy in human–agent interactions, the following research questions arise:

- How to understand empathy in order to assess empathy in human–agent interactions?
- What are the minimum necessary conditions for assessing computational empathy?

Through a systematic literature review of prior research on empathy in CAs, data was collected regarding the notion of empathy in these systems. Proceeding the data collection, a set of qualitative data modeling steps was conducted following a Grounded Theory (GT) methodology (Chun Tie, Birks, & Francis, 2019); The collected data was analyzed and conceptualized, resulting in an ontology for computational empathy, which we implement using Web Ontology Language (OWL).[1] This is the first computational semantic model, introduced as an ontology, in the state-of-the-art that captures functional necessary conditions for computational empathy.

We present the potential of the approach by applying it as a qualitative tool for assessing empathy in interactive software agents in the area of health and well-being. The evaluation is based on a controlled user-study, where participants have interacted with two of the most popular health and well-being chatbots (Wasil, Palermo, Lorenzo-Luaces, & DeRubeis, 2021), Replika[2] and Wysa,[3] and answered a proceeding assessment protocol based on the proposed definition of empathy. See Fig. 1 for an overview of the conducted research methodology. To the best of our knowledge, no previous effort has been done to assess computational empathy as introduced in the current work. Findings show that our definition can distinguish empathy in CAs on different levels of understanding by looking at different sub-dimensions of empathy.

The main contributions of the paper are: (1) a formal multi-dimensional definition for computational empathy, (2) a qualitative methodology for evaluating computational empathy, specifying an assessment protocol and an analysis method using precise and imprecise uncertainty measures, (3) an OWL ontology implementation, provided as open access material, and (4) a qualitative assessment of computational empathy of two state-of-the-art chatbots, Replika and Wysa, by considering precise and imprecise uncertainty measures.

This paper is organized as follows. Section 2 presents seminal and recent work for defining and measuring empathy. In Section 3, a literature review of prior work in CA–human interaction is presented, on which we inform our definition of empathy. In Section 4, the data analysis and qualitative data modeling steps are presented. In Section 5, our formal definition of empathy, comprising six key dimensions, is presented. In Section 6, an evaluation of the definition is presented, applying it as a tool for measuring empathy in CAs. Sections 7–8 concludes the paper with a discussion regarding limitations, strengths, possible applications, and directions for future work.

2. Related work

Empathy has been understood as a multi-dimensional process that can be divided into two overall aspects, (1) affective/mirroring processes and (2) cognitive/reconstructive processes (Goldman, 2011). The affective processes concern arousal from, and (involuntary) response to, an agent's expressed experiences, and the cognitive processes concern deliberation and interpretation of an agent's experiences and delivering an inferred understanding back to the agent through interactions (Elliott et al., 2018). In this direction, empathy has been suggested as a complex interaction between affective and cognitive processes, in terms of inversely related influences between (1) theory of mind constructions, (2) awareness of affective responses, and (3) primitive affective resonance induction (Watt, 2005). From a social cognition viewpoint (Szanto & Krueger, 2019), empathy has been considered in relation to interconnected and shared emotions. In this view, interpersonal social relations and groups with shared emotions are considered

[1] Source of OWL ontology: https://github.com/ComputationalEmpathy/empathy-ontology.

[2] Chatbot Replika: https://replika.ai.

[3] Chatbot Wysa: https://wysa.io.

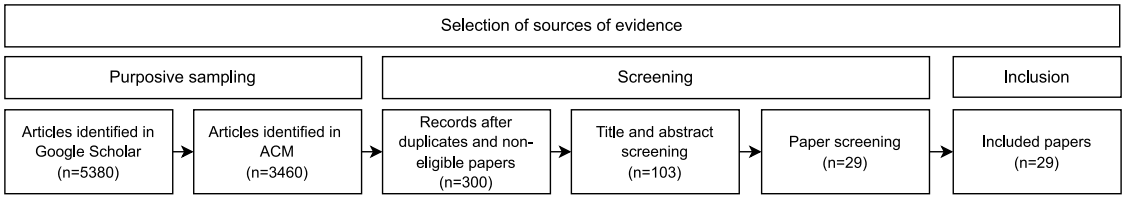


Fig. 2. Systematic literature review, following guidelines provided by the PRISMA framework (Rethlefsen et al., 2021).

as important social constructs related to the definition of empathy. From a communication theory viewpoint, the Interpersonal Adaptation Theory (IAT) (Burgoon, Stern, & Dillman, 1995), understood to be closely related with empathy, addresses ways in which individuals adapt to one another in interactions to, for instance, mimic and synchronize behavior, social relation and personality with another. This highlights an interpersonal aspect of empathic capability.

In the literature, different measures and models of human empathy have been developed. Let us look at some of them next.

2.1. Measures of empathy

A variety of measures have been informed by previous definitions, used as tools for assessing empathy in humans. These measures have typically relied on self-reports, for instance the Questionnaire Measure of Emotional Empathy (QMEE) (Mehrabian & Epstein, 1972) and Hogan Empathy Scale (HES) (Hogan, 1969), mostly focusing on affective/emotional empathy, leaving out cognitive empathy, or vice versa. A more recent work by Jolliffe and Farrington (2006) (Jolliffe & Farrington, 2006) criticizes the one sided approaches of QMEE and HES. Jolliffe and Farrington further criticize that QMEE and HES only capture hypothetical scenarios, with questions in the format "I feel [hypothetical emotion] when I see [hypothetical scenario]". Instead, they propose a measure focusing on a participant's previous experiences, and taking into account both cognitive and affective empathy, the so-called Basic Empathy Scale (BES) (Jolliffe & Farrington, 2006), with questions based on four 'basic emotions' (fear, sadness, anger and happiness). Examples of questions in BES are "I get caught up in other people's feelings easily" (affective) and "I can often understand how people are feeling even before they tell me" (cognitive) (Jolliffe & Farrington, 2006).

While these previous measures of empathy have been focused on self-reports, to measure empathy of the individual, we can observe limitations in these measures to capture empathy in interactive software agents, because (1) they measure the individual's own empathy, in comparison to an interlocutor's perceived empathy, and (2) a software agent may not have the same criteria for empathy as humans. In order to measure computational empathy, we need a formal definition of what empathy means in a software agent, and if that measure should be derived from a human's perspective, then we need an assessment protocol with questions focused on an individual's perceptions of an agent's empathic capabilities in the interaction. These questions must be based on aspects that can be perceived and understood by the participant, and the formal model must be able to infer higher level concepts based on low level perceptions.

2.2. Models of empathy

A variety of models of empathy have been developed, used to better understand empathy in humans in a broad sense. For instance, the Perception and Action Model (PAM) (Preston, 2007). PAM is a model for empathy developed in cognitive neuroscience, that defines empathy as a shared emotional experience occurring when an individual comes to feel a similar emotion to another as a result of perceiving, or imagining, the other's (emotional) state. PAM is informed by findings that there are shared representations (activation) in the human brain for perceiving and generating action. For instance, as an individual perceives a person crying, similar brain activation can be measured as if the perceiver was crying. PAM is further extended, such that perception can arise from situations where the subject is not directly perceiving an object, but imagines the state of the object (in the current work, we refer to these kind of phenomenons of an agent's mental state as its *manifestation*). Given the neuroscience perspective of PAM, we see limitations in using this theory alone for modeling computational empathy in interactive software agents, but it is an important theoretical base to be acknowledged; When modeling cues of empathy in an interlocutor agent, dependencies between the agent's perception and action, described by PAM, are critical to be captured by the model.

A recently defined model of empathy is the Nine Dimensions of the Empathic System (ES9) (Guthridge & Giummarra, 2021). ES9 considers empathy as a complex series of multi-dimensional processes that comprises an "empathic system". In their taxonomy, empathy is defined in terms of nine high level dimensions, (1) catalyst, (2) function, (3) process, (4) outcome, (5) affective state, (6) cognitive state, (7) self and other, (8) behavior and (9) context. This is relevant for getting an overarching perspective on human empathy. A limitation in ES9 is that the definition consists of abstract high-level concepts, not intended to be measurable. To enhance precision, abstract notions like "affective state" (in ES9) need concrete, observable, low-level counterparts for objective assessment. Furthermore, ES9 lacks computational applicability, requiring further exploration and modeling to be integrated in automated reasoning. This highlights the need for a formal definition and computational model of empathy.

The following section presents a literature study of prior work in CA-human interaction, analyzed through a Grounded Theory process (Chun Tie et al., 2019), on which we inform our definition of computational empathy.

3. Systematic literature review

We conducted a systematic literature review to gather data on conversational agents (CAs) in the context of health and well-being, specifically focusing on the concept of empathy within these systems (refer to Figure Fig. 2). Adhering to the PRISMA guidelines (Rethlefsen et al., 2021) for systematic literature reviews, our search yielded numerous results that aligned with our predefined criteria. Subsequently, an examination of titles and abstracts against the inclusion criteria led to the final inclusion of 29 papers for further data analysis (see Table 2). We now outline the methodological steps of the literature review.

Search Strategy. The search strategy regards identifying relevant literature to be further assessed for eligibility and inclusion. Due to the available literature on chatbots varies in terms of approaches (e.g., technical and non-technical perspectives), a decision was to use general research databases containing both conceptual and technical perspectives. Data were collected through two databases; Google Scholar and ACM Digital. We did not make use of cited references or citing references of included papers, nor including additional studies from other information sources, focusing only on the original papers based on our search criteria. The main search term of ["Conversational agent" OR "Chatbot"] was used due to common CA applications in the

Table 1
Relevance assessment criteria.

Quality criteria	Score
Do authors clearly state the purpose of the chatbot(s)?	Yes = 1; No = 0
Do authors address psychological theories?	Yes = 1; No = 0
Do authors define empathy?	Yes = 2; No = 0
Do authors evaluate chatbot(s)?	Yes = 1; No = 0
Are there any considerations on trustworthiness?	Yes = 1; No = 0
Do authors clearly state limitations?	Yes = 1; No = 0

area of health and well-being. Additional search terms used were AND ["Well-being" OR "Health"] AND ["Empathy"]. In Fig. 2, we illustrate the systematic review process guided by specified eligibility criteria: (1) Publication date between 1950–2021, (2) Papers need to be in English, (3) Papers need to be peer-reviewed, and (4) Papers need to include a definition of empathy, an evaluation of empathy, and describe an interactive system in health and well-being.

Relevance Assessment Criteria. The papers that met the inclusion criteria were assessed by using a relevance assessment criteria (see Table 1), similarly to Ahmad et al. (2018). Considering our key focus on defining empathy, we found it crucial to search for papers that defined empathy from an empirical and theoretical stance. Following this train of thought, we aimed to include evaluations of chatbots in terms of an indirect assessment of a definition of empathy. Our criteria were also informed by the vast research body of empathy from perspectives outside computer science, e.g., psychology. Notably, trust and empathy are closely connected (Gebhard et al., 2021), leading us to include author considerations on trustworthiness. Lastly, we included limitations as a relevance criteria to ensure rigor of included papers. Over and above established inclusion criteria, this quality tool helped us to structurally assess essential aspects of the papers. The included papers had to score at least '2' on the relevance assessment criteria (Table 1) to be considered for final selection. Giving a paper a score of 2 was primarily reserved for papers that provided a clear and specific definition of empathy, as this aspect was considered the most crucial for the study. Other factors within the criteria were assigned a score of 1 each, indicating their importance in adding relevance to the research.

Data Extraction Process. The included papers were carefully assessed and re-read by multiple authors. In a structured way, following established criteria, we discussed and extracted a notion of empathy for each paper. As shown in Table 2, empathy notions take shape in different ways, e.g., humor, relations, or sociability. The included papers met our inclusion and relevance criteria by displaying different notions of empathy. We utilized the five Ws (El-Masri & Mowbray, 2019), asking who, what, when, where, and why different notions of empathy play a role. We extracted relevant sections of the material as a final preparatory step of the Systematic literature review, before employing the first step of the Grounded Theory process.

4. Data analysis and modeling

Proceeding the data collection, a series of qualitative data modeling steps was conducted following the methodology of Grounded Theory (Chun Tie et al., 2019). Grounded theory is a well established method for knowledge engineering (Lifschitz, Porter, & Van Harmelen, 2008; Pidgeon, Turner, & Blockley, 1991; Rose-Davis, Van Woensel, Abidi, Stringer, & Abidi, 2022; Yuen & Richards, 1993; Zhang, Wang, & Li, 2023), used to code, categorize and relate qualitative data. Grounded theory gives a clear methodology for knowledge elicitation for building qualitative knowledge bases. This iterative method supports traceability and explainability in the elicitation of knowledge. In this way, we can follow the knowledge engineering process from collected qualitative data to the final computational model. Grounded theory is particularly well-suited for our study due to its inherent capacity to discover and define concepts in terms of their relationships, a quality that resonates with our objective of establishing a structured model of computational empathy.

The grounded theory process includes the following steps; (1) initial coding, identifying segments of information in the raw data extracted from the selected papers, (2) intermediate coding, clustering the initial codes into categories, and (3) advanced coding, defining overall themes by clustering the categories of the previous step. This is a process of finding recurring concepts in the qualitative data that collectively explain the sought after, more abstract, concepts.

Theoretical saturation: The aim of the data analysis is to find a minimal set of themes (which we refer to as dimensions) constituting a comprehensive platform for computational empathy. An important benchmark when aiming for a minimal set of concepts from qualitative data is the notion of *theoretical saturation* (Saunders et al., 2018), referring to a point in the data analysis when no additional insights or concepts emerge from data and a set of overall themes have been defined, backed up by a hierarchy of codes and categories originating from the data (see Table 3). While there are no explicit rules for deciding when theoretical saturation has been reached, there are methods for making such approximations. Saturation is understood as a degree of completeness, where further data-collection results in increased diminishing returns, i.e., any new code or category is either a synonym or can be subsumed in previously defined concepts. It is argued that theoretical saturation is reached when "sufficient depth of understanding" (Saunders et al., 2018) has been established in relation to emergent categories and themes. In the current study, "depth of understanding" is understood as levels in a graph. The graph starts at an abstract root concept which we call "Computational empathy". Each descending level corresponds to less abstract concepts, where each node is a partial explanation of its parent node in the graph. Theoretical saturation was declared when (1) the lowest level concepts could be objectively understood, and (2) the parent nodes, i.e., categories and themes, could be explained based on their child nodes, and (3) the overall themes covered the root concept on an abstract level. In order to provide an understanding of the completeness of the model, we analyzed it by considering empathy aspects from a selection of cognitive theories. In particular, it was analyzed by considering PAM (Preston, 2007), Social cognition (Bae Brandtzæg et al., 2021) and IAT (Burgoon et al., 1995), previously described in Section 2. This process resulted in a set of 6 overall themes, in a structure comprised of three levels of understanding.

In what follows, we describe the three steps of our data analysis, finally resulting in a graph structure on which we base the proposed formal model of computational empathy.

Initial coding. This step aimed to capture the substance of the data and break it down into smaller segments, referred to as codes. This was done by highlighting the data that were in line with the initial search queries. If a pattern was found in the codes, such that they were frequently found in multiple papers, then these were combined into dominant codes. Initially, 80 codes were generated which regarded meaningful units focusing on the definition of empathy. Hence, given the selected raw material, codes were extracted to aggregate the qualitative information. For instance, as done in the following data segments: *"[...] expected the chatbot to retain context [CONTEXT-SENSITIVE] across chat sessions, thus providing users with personalized [PERSONALIZED] recommendations [...]"* (Jain et al. p. 901), *"sentiment [SENTIMENT] and emotion detection [EMOTION DETECTION] technique may be utilized to identify the corresponding [...]"* (Rahman et al. p. 13); *"An empathic human-like [HUMAN-LIKE] chatbot avatar promoting natural conversations can promote user engagement [ENGAGING] towards chatbot services."* (Chen et al. p. 222). This aggregation step is further presented in Table 4. The next step is to cluster the codes into categories, further aggregating the information while preserving their links to the qualitative material.

Table 2
Included papers which underwent further data collection.

Paper	Year	Chatbot context	Empathy notion
Bae Brandtzæg, Skjuve, Kristoffer Dysthe, and Følstad (2021)	2021	Social support	Emotional Support
Beilharz et al. (2021)	2021	Supporting disorders	Therapeutic bonds; building rapport
Cameron et al. (2018)	2018	Self-assessment mental health	Error management; personality and understanding
Casas et al. (2021)	2021	Empathic social chatbot	Detect emotions and respond accordingly
Ceha, Lee, Nilsen, Goh, and Law (2021)	2021	Social well-being	Humor; self-defeating jokes evoke empathy
Chen, Lu, Nieminen, and Lucero (2020)	2020	Social support	Engaging; human-like
Chung, Cho, and Park (2021)	2021	Obstetric and mental health	Sympathetic responses; facilitate sharing knowledge
Croes and Antheunis (2021)	2021	Social companion	Goal-oriented; contextual; timing; good listener
De Gennaro, Krumhuber, and Lucas (2020)	2020	Social ostracism	Show understanding; empathic feedback
Gabrielli, Rizzi, Carbone, and Donisi (2020)	2020	Self-help mHealth	Reply emphatically; social awareness; interpersonal relationship
Ghandeharioun, McDuff, Czerwinski, and Rowan (2019)	2019	Behavior change	Emotionally appropriate; acknowledge users emotional state
Greer et al. (2019)	2019	Well-being support	Notice and acknowledge attainable goals
Grové (2021)	2021	Encouragement	Personality; adaptable to mood
Hauser-Ulrich, Künzli, Meier-Peterhans, and Kowatsch (2020)	2020	Build alliance	Affective empathy; supportive; coaching
Inkster, Sarda, and Subramanian (2018)	2018	Mental resilience	Balanced engagement efficiency
Jain, Kumar, Kota, and Patel (2018)	2018	Social, chit-chat	Sensitive to context; personalized
Jang et al. (2021)	2021	Increase concentration	Empathic responding; build relationship
Kraus, Seldschopf, and Minker (2021)	2021	Trust in HCI	Understand statements; behaviors or feelings of another
Lee et al. (2019)	2019	Self-care	Relate to suffering; Self-compassion
Lee, Yamashita, and Huang (2020)	2020	Promote self-disclosure	Understand people; tone-aware
Li et al. (2021)	2021	Empathetic chatbot	Understand feelings and experiences; emotion causes
Lin et al. (2019)	2019	Emotion detection; empathic responses	Express and perceive emotion; social bonding
Ly, Ly, and Andersson (2017)	2017	Self-reflection, practice behaviors	Tailored mood; tailored content
Maeda et al. (2020)	2020	Optimize preconception health	Humanness; acknowledge emotions
Medeiros, Gerritsen, and Bosse (2019)	2019	Emotional support	Expressing empathy; supportive messages
Morris, Kouddous, Kshirsagar, and Schueller (2018)	2018	Empathic support	Perceive causes to emotions
Rahman et al. (2021)	2021	Health advice	Sentiment; emotion detection
Ryu et al. (2020)	2020	Depression support	Sociability; engaging in interactions
Sia, Yu, Daliva, Montenegro, and Ong (2021)	2021	Student's well-being	Show understanding; show affect; facilitate sharing

Intermediate coding. At this stage, 77 codes were clustered in terms of their relation to each other. Thereafter, a second process of naming the clusters was conducted. This resulted in 13 labeled categories: *Traits, Acts, Manifestations, Expressions, Moves, Activities, Interpersonal, Building trust, Functional performance, Context, Emotions, Mentalization, and Perception*, each defined as a set of codes (see Table 3). The next step is to cluster the categories into an overarching set of themes.

Advanced coding. In an iterative way, categories were further analyzed, by clustering and revising, adding or removing codes to fit the characterization of a dimension of empathy. This resulted in a set of 6 refined dimensions: {*Perceive, Act, Theory of mind, Manifest, Ratification, and Interpersonal*}. The dimensions are further related to underlying categories and initial codes, forming a hierarchical graph structure. By defining the themes following the results of the previous steps, we can preserve their links to the initial qualitative material. By reaching theoretical saturation in each step, we show that these themes are minimal, referred to as the 6 necessary conditions of computational empathy.

4.1. Grounded theory outcome

The grounded theory analysis steps resulted in a hierarchy of codes, categories and themes. A final step, which is important for the proceeding computational modeling process, is a cleaning step, where codes and categories are further refined. This resulted in a set of concepts that jointly define a taxonomy for computational empathy. The highest level of the taxonomy consists of the following set of 6 main concepts: {Perceive, Theory of mind, Act, Manifest, Ratification, Interpersonal}, each being a set of sets of sub-components. Let us explain the intended meaning behind each of the main concepts and their components.

Perceive = {*Emotion, Behavior, Mood, Message* = {*Body language, Facial expressions, Presence, Speech, Text*}, *Interaction property* = {*Response rate, Time between sessions, Duration, Response content level*}}. The intended meaning of *Perceive* is the following description: The ability to perceive various concepts from external stimuli is crucial for establishing specific aspects of empathy and theory of mind. Depending on the nature of agents and the mode of interaction, diverse perceptual abilities become relevant for empathy. In certain human-agent interaction contexts, the perception of mood, emotion, behavior, and communicative messages holds significance. Moreover, as interactions extend into the long term—spanning multiple transactions or sessions—a range of interaction properties gain relevance. This includes the capacity to perceive the other agent's response rate, time between sessions, duration of interaction, and the level/extent of response content.

Theory of Mind, ToM = {*Agreements, Emotions, Goals, Mood, Needs, Personality, Rapport, Social bond, Trust, Values, Errors* = {*Deception, Misunderstanding, Misperception*}}. The intended meaning of *Theory of Mind* is the following description: Understanding what is going on in the mind of the other is a key aspect of empathy. This understanding is built upon the available information and knowledge—what is perceived or previously known. These assessments may encompass various aspects of the other agent's mental state, including mood, emotions, and relational dynamics such as rapport, trust, and social bonds. Deeper insights into the other agent's values, goals, and needs may require more

Table 3
Data aggregation process. This table presents how codes are clustered into categories and dimensions.

Initial coding: Codes (n=77)
Emotional self-awareness CAT [11], Perspective-taking CAT[7,8,12], Social Awareness CAT[7,11], Sense, hurt or pleasure as he senses it CAT[7,8,12], Human-like CAT[3,7,12], Curious CAT [7], Naturally understand CAT[7,8,12], Tone-aware CAT[9,10], Suffering CAT[2,4,5,12], Self-defeating CAT[1,2,4,5,8], Natural CAT [7], Savoring CAT[2,5], Lack of humanity CAT[3,7,12], Sadness/loneliness CAT[1,2], Appreciation CAT [3], Positive emotion CAT [11], Intelligence CAT[1,2,10], Personalized CAT[1,2,5,7,10], Humour CAT[1,2,5,6], Mood CAT[4,2,10], Inspiring CAT[1,2], Charismatic CAT[1,2], Affective CAT [4], Engagement CAT[7, 8], Conflict resolution CAT [6], Comforting CAT[5,8], Coaching CAT [6], Engaging CAT[4,6,7,8,10], Goal-oriented CAT[4,6], Tailored CAT[5,10], Positive reappraisal CAT [6], Good listener CAT[2,5,7,10], Acts of kindness CAT[2,4,5,7], Build rapport CAT[7,8,10], Curiosity CAT [7], Grateful CAT[2,5], Sociability CAT[1,2,7], Humanness CAT[3,7,12], Acknowledge CAT[3,4], Interpersonal relationship CAT[2,7,8], Emotionally appropriate CAT[4,8,9,11], Detect emotional states CAT [13], Context-sensitive CAT[5,10,13], Emotion detection CAT [13], Contextual CAT [5], Relationship CAT [7], Fun CAT[1,2], Friendly CAT[1,2,5,7,8], Adaptable to mood CAT[9,10], Show understanding of others emotional state CAT[3,4,11], Concern feelings and experience of others CAT[7,8,11,12], Show understanding CAT[3,4], Show affect CAT [3], "feel for" CAT[3,4], Acknowledge others feelings CAT[3,4], Emotionally expressive CAT[3,4,11], Supportive CAT[4,6,8], Sentiment CAT[3,4], Expressively emotionally supportive CAT[3,4,11], Express/Perceive emotion CAT[3,4,11], Support CAT[4,8], Notice CAT[3,4], Sympathy CAT[2,4,12], Understand/respond to emotions appropriately CAT[4,8,9,11], Understand feelings and experience of others CAT[7,8], Understand statements, behaviors and feelings of another CAT[7,8], Accurately perceive frame of reference of other CAT[4,8,9], Relate and respond accordingly to emotions CAT[4,8,9,11], Social bonding CAT[1,2,6,7], Understanding CAT[12,13], Humorous CAT[1,2], Intelligent CAT[1,2], Personality CAT[1,2], Persuasive CAT [2], Strategic CAT [3], Avoid negative words CAT [8]
Intermediate coding: Categories (CAT, n=13)
1:Empathic traits DIM[A], 2:Empathic acts DIM[A,M], 3:Empathic manifestation DIM[M], 4:Empathic expressions DIM[A,M], 5:Empathic moves/skills DIM[A], 6:Empathic activities DIM[A], 7:Interpersonal DIM[I], 8:Building trust DIM[R], 9:Functional performance DIM[P,T,A,M,R,I], 10:Context DIM[P], 11:Emotions DIM[P,T], 12:Mentalization DIM[T], 13:Perception DIM[P,M]
Advanced coding: Dimensions of Empathy (DIM, n=6)
P:Perceive, A:Act, T:ToM, M:Manifest, R:Ratification, I:Interpersonal

sophisticated reasoning. Additionally, integral components of empathy involve error detection, aiming to identify misperceptions or misunderstandings, and the ability to discern deception, such as untruthful statements, from the other agent. This dimension relates to cognitive/reconstructive processes in human empathy theories (Goldman, 2011).

Act = {*Emotion, Behavior, Mood, Context, Message* = {*Body language, Facial expressions, Presence, Speech, Text*}, *Interaction property* = {*Response rate, Time between sessions, Duration, Response content level*}}. The intended meaning of *Act* is the following description: The capacity to generate empathic responses that are perceived by the interacting agent, shaping and influencing social bonds, constitutes a key aspect of empathy. These responses are rooted in the agent's perception of the situation and its predictive abilities regarding the other's mental state (Theory of Mind). Acting to demonstrate understanding and express affect becomes crucial for the agent. In certain human-agent interaction contexts, actions aimed at sharing the agent's mood, emotion, behavior, and communicative messages hold significance. Moreover, as interactions extend into the long term—spanning multiple transactions or sessions—a range of interaction properties gain relevance. Properties such as response rate, time between sessions, duration of interaction, and the level/extent of response content become relevant.

Manifest = {*Agreements, Emotions, Goals, Mood, Needs, Personality, Rapport, Social bond, Trust, Values, Errors* = {*Deception, Misunderstanding, Misperception*}}. The intended meaning of *Manifest* is the following description: Being influenced by the perceived situation of other agents, such as their moods and emotions, leading to a transformation in its internal state or manifestation is a key aspect of empathy. This internal manifestation is central to the process of mentalizing and acquiring a Theory of Mind of the agent, which, in turn, influences its actions and shares its internal state. As manifestations unfold, the agent begins to shape a personality in relation to the other agent. The capability to internalize the situation of another agent fosters interpersonal interactions, wherein the agents mutually impact each other's mental states and behaviors throughout the course of their interaction. This dimension relates to affective/mirroring processes in human empathy theories (Goldman, 2011).

Ratification = {Acceptance, Agreements, Equality, Mutuality}. The intended meaning of *Ratification* is the following description: Ratification empowers agents to attain a state of acceptance and agreement. Ratification further comprises two crucial components—equality and mutuality. Equality emphasizes seamless interaction, ensuring each agent has ample opportunities for action and interpretation of incoming actions. Mutuality embodies positive communication, fostering a deep understanding between agents while minimizing misunderstandings, misperceptions, and deception. The synergy of equality and mutuality facilitates consensus-building and agreement between agents. Ratification, in essence, nurtures tolerance toward the other agent, promoting forgiveness and enhancing activities that build interpersonal trust.

Interpersonal = {*Social bonding, Tolerance, Trust building, Interconnection* = {*Connected emotion, Connected mood, Connected behavior, Connected message*}}. The intended meaning of *Interpersonal* is the following description: In empathy, a fundamental aspect revolves around the interaction that builds interpersonal trust. Throughout such interactions, agents impact each other's behaviors and internal manifestations. In human-agent interactions, interpersonal connections prove crucial for perceived empathy, facilitating the sharing and mutual influence of mood, emotion, behavior, and knowledge. These interconnected states play an important role in cultivating and sustaining trust, tolerance, and social bonds. Establishing interpersonal trust depends on factors like consistency, honesty, openness, and the steadfast honoring of commitments.

In this section, we have described how our employed methodological steps inform a multi-dimensional understanding of empathy. The empirical understanding of the dimensions of empathy characterizes our formal definition of computational empathy, as described in the following section.

5. Formal definition of empathy

Following the grounded theory results, the identified empathy dimensions are formalized in terms of an ontology, defining the concept we call "computational empathy". The computational empathy ontology is a representation, and we expect perceptions of an agent's empathy to be represented there. By aggregating the perceptions through the ontology, we get *performance measures* of an agent in terms of computational empathy dimensions.

The ontology is organized into a taxonomy, where general and abstract concepts are positioned at a higher level, complemented by more specialized and less abstract concepts at a lower level. This structure creates a hierarchy of classes, outlining conditions for empathy across different levels of understanding. At the first level, six minimal necessary conditions for empathy are defined. Progressing to the second

Table 4
Extracted coding definitions. This table presents a subset of codes resulting from the Initial Coding of the collected data.

No.	Coded definitions	Source
1	"A high performing ML model would become a necessity when conversation volumes increase to ensure high user engagement [ENGAGING] and retention [RELATIONAL]."	Inkster et al. p. 9
2	"understanding how our feelings and emotions work. [EMOTIONAL SELF-AWARENESS]"	Gabrielli et al. p. 3
3	"[...] understand the statements, behaviors or feelings of another person, from the counterpart's perspective or preconditions. [EXPRESS/PERCEIVE EMOTIONS]"	Kraus et al. p. 358
4	"providing an emotionally appropriate response [...], similar to what happens in a successful human-human interaction [...] [EMOTIONALLY APPROPRIATE]"	Ghandeharioun et al. p. 13
5	"the empathizing agent communicates his or her understanding of the other individual's emotional state [SHOW UNDERSTANDING OF OTHERS EMOTIONAL STATE]"	de Gennaro et al. p. 3
6	"[...] users' perception of Abot's human-like qualities and affective abilities and acceptability as a chat companion [SHOW UNDERSTANDING]"	Sia et al. p. 39
7	"Each message sequence begins with a warm greeting, in which the chatbot enquires about the participant's mood and replies in an empathic way [SHOW AFFECT]"	Hauser-Ulrich et al. p. 5
8	"However, agents with sophisticated empathic abilities (i.e., agents that seem to truly understand the user's emotional experience) [ACCURATELY PERCEIVE FRAME OF REFERENCE OF OTHER]"	Morris et al. p. 2
9	"[...] expected the chatbot to retain context [CONTEXT-SENSITIVE] across chat sessions, thus providing users with personalized [PERSONALIZED] recommendations [...]"	Jain et al. p. 901
10	"sentiment [SENTIMENT] and emotion detection [EMOTION DETECTION] technique may be utilized to identify the corresponding [...]"	Rahman et al. p. 13
11	"An empathic human-like [HUMAN-LIKE] chatbot avatar promoting natural conversations can promote user engagement [ENGAGING] towards chatbot services."	Chen et al. p. 222
12	"empathy is the ability to understand[UNDERSTAND] and concern the feelings and experience of others [CONCERN FEELINGS AND EXPERIENCES OF OTHERS"	Li et al. p. 2041
13	"Emotional support - Expressions of empathy [EMOTIONALLY EXPRESSIVE], love, trust, and caring [CURIOUS]"	Brandtzaeg et al. p. 6
14	"older users dealing with anxiety could potentially benefit from spare-time interactions with technologies that provide engaging interactions. [SOCIABILITY] [SOCIAL BONDING]"	Ryu et al. p. 20
15	"Compassion and empathy are associated, but are not the same ["FEEL FOR"]. Empathy allows people to relate to others suffering [SUFFERING] cognitively and affectively"	Lee et al. p. 8
16	"[...] giving compassion to Vincent (or another being) than towards oneself may be more natural [NATURAL, NATURALLY UNDERSTAND, TONE-AWARE] in conversational contexts"	Lee et al. p. 8
17	"Self-defeating [SELF-DEFEATING] humour [HUMOUR] is characterized by an excessive use of self disparaging humour, by which the user attempts to amuse others at their own expense."	Ceha et al. p. 4
18	"the ability to detect another person's emotional state. [EMOTION DETECTION] [...] relate to these emotions and respond according to how the other person must be feeling [RESPOND AND RELATE ACCORDINGLY]"	Casas et al. p. 2
19	"asked participants to what extent they agreed with these statements: "Mitsuku said the right thing to make me feel better,"[CONTEXTUAL] "Mitsuku responded appropriately to my feelings and emotions," "Mitsuku came across as empathic," "Mitsuku said the right thing at the right time," [TIMING]] and "Mitsuku was a good listener."[GOOD LISTENER] [GOAL-ORIENTED]"	Croes et al. 287
20	"Empathetic chatbots are conversational agents that can understand user emotions and respond appropriately, which is an essential step toward human- like conversation. " [UNDERSTAND/RESPOND TO EMOTIONS APPROPRIATELY] "[...] humans express and perceive emotion [EXPRESS/PERCEIVE EMOTION in natural language to increase their sense of social bonding. [SOCIAL BONDING]"	Lin et al. p. 1
21	"as empathic responses based on user's mood [MOOD], tailored [TAILORED] content based on user's previous inputs, daily check-ins to create a sense of accountability, and weekly summaries in the end of each week containing [...]"	Ly et al. p. 42
22	"express empathy to humans in a way that they perceive [PERCEP] it is natural.[NATURAL] [NATURALLY UNDERSTAND]"	Medeiros et al. p. 234
23	"Because as weird as it is talking to a robot [SAVORING], it is nice to vent and be able to see [NOTICE] others with cancer talking and speaking out [ACKNOWLEDGE] about how they coped or felt during their treatment. Seeing that I'm not alone and having someone guide me to find the positives [ACTS OF KINDNESS] in my life now is really helpful. [GRATEFUL]"	Greer et al. p. 6
24	"Ash is the guide, it is inspiring and charismatic [CHARISMATIC]], fun [FUN], friendly [FRIENDLY], empathic, humorous [HUMOUR] [PERSONALITY] [MOOD] [ADAPTABLE TO MOOD]"	Grové et al. p. 7
25	"The 7 categories are: Answering, Error management, Intelligence, Navigation, Onboarding, Personality [PERSONALITY]] and Understanding [UNDERSTANDING] [INTELLIGENCE]"	Cameron et al. p. 126
26	"A lack of humanity [LACK OF HUMANITY] or empathy, e.g. 'robotism', 'coldness'[COLDNESS]] or 'one-way interaction') was mentioned"	Maeda et al. p. 1139
27	"the use of a pictorial character within conversation text appeared to be a useful "shortcut" for building rapport [BUILD RAPPORT] with users."	Beilharz et al. p. 9
28	"Todaki uses dialogue skills such as empathic responding and reflection [APPRECIATION] [CURIOSITY] to create a therapeutic relationship with the users. [RELATIONSHIP] [INTERPERSONAL RELATION]"	Jang et al. p. 4
29	"explained the need for partner support in a friendly tone [AVOID NEGATIVE WORDS][SUPPORTIVE] and delivered practical strategies with relevant images in which a man actively supported [SUPPORT] his partner, showing empathic concerns and sympathetic responses"	Chung et al. p. 6

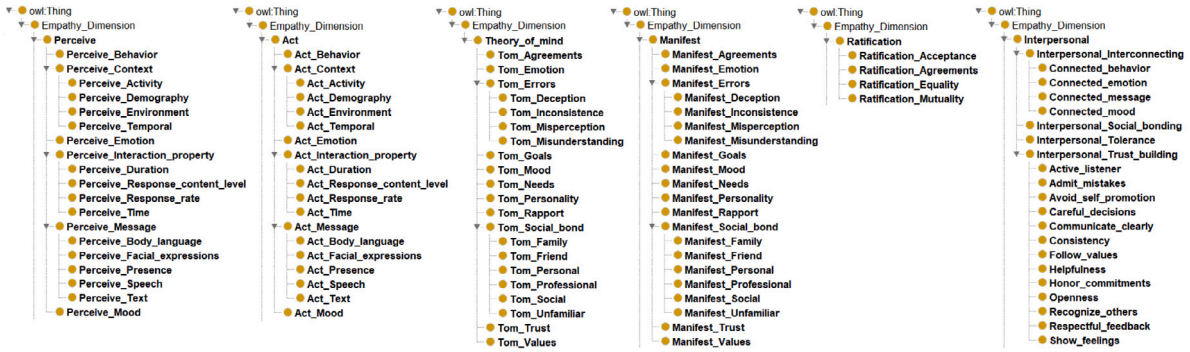


Fig. 3. Empathy ontology implemented in OWL format.

level, there are application-specific dimensions, each holding varying relevance in specific settings. The third level further refines these dimensions, incorporating specialized and more directly or indirectly measurable concepts. These concepts can be nuanced based on specific theories, such as cognitive theories. In this proposed ontology, while we consider the first level to outline the necessary conditions, the second and third levels of the taxonomy introduce sub-concepts. It is important to note that these sub-concepts are not universally necessary; their relevance may vary across different contexts and for different agents.

In order to enable interoperability with a wide range of knowledge-based systems, such as on the semantic web, the ontology is implemented in Web Ontology Language (OWL) (see Fig. 3). This enables integration of external ontologies to define sub-concepts of computational empathy. This is a motivation for maintaining a hierarchical structure in the ontology, making intended meanings of concepts less abstract on a low level, supporting conceptual alignment (Kolli, 2023; de Souza & Davis, 2004; Zimmermann & Le Duc, 2008) with other ontologies.

5.1. The description logic $\mathcal{ALC}$

The proposed ontology is expressed following the syntax of the description logic $\mathcal{ALC}$ (Attribute Language with general Complement) (Baader, Calvanese, McGuinness, Patel-Schneider, et al., 2003) (see Table 5). Description Logics (DLs), such as $\mathcal{ALC}$, are formal languages for expressing ontologies—conceptual representations of domain knowledge that are both human-readable and machine-processable for automated reasoning. DLs involve defining concepts (classes), individuals (objects), and roles (properties) together with logical operators (e.g., negation, union and intersection). A terminology is established to define concepts and relationships (TBox). Then, a knowledge representation system can be employed to store and reason about this terminology using assertions about individuals (ABox). Together, the TBox and ABox constitute a so-called ontology, a structured and computational method for representing and reasoning about qualitative knowledge (Baader, Calvanese, McGuinness, Nardi, & Patel-Schneider, 2003). In the syntax of $\mathcal{ALC}$, let A be an atomic concept, C and D be concepts, and R be a role. Then, the syntax of $\mathcal{ALC}$ enables the definition of TBox axioms, such as $C \sqsubseteq D$, indicating that concept C is a sub-concept of D. It also enables the expression of ABox assertions, such as $C(a)$, signifying individual a belonging to concept C, and $R(a, b)$, signifying the relation between individuals a and b through role R. Beyond knowledge representation, $\mathcal{ALC}$ provides robust reasoning capabilities, including the inference of implicit knowledge, assessment of concept satisfiability, and detection of contradictions. Notably, it supports entailment, assessing the truth of ABox assertions or TBox axioms across all knowledge base models, as well as conjunctive query entailment for variable-containing queries and the combination of concept and role expressions. For instance, in $\mathcal{ALC}$ notation, a conjunctive query can be represented as $C(x) \sqcap D(x)$, seeking individuals x that simultaneously belongs to Concept C and Concept D. When this query is evaluated against provided TBox axioms and ABox assertions, it will return individuals that satisfy these conditions.

Table 5
Syntax of the description logic $\mathcal{ALC}$. A is an atomic concept, C and D are concepts, and R is a role.

Syntax		Description
$C, D ::=$	A	atomic concept
	$\top$	universal concept
	$\bot$	bottom concept
	$\neg C$	negation
	$C \sqcup D$	union
	$C \sqcap D$	intersection
	$C \sqsubseteq D$	inclusion
	$C \equiv D$	equivalence
	$\exists R.C$	existential restriction
	$\forall R.C$	universal restriction

5.2. Taxonomy of computational empathy

Given the overall concept of $Computational\_Empathy$ (referred to as Level 0), comprised of parts, so-called $Empathy\_dimension \sqsubseteq Computational\_Empathy$, the model states that the minimum necessary conditions for computational empathy is a set of 6 main dimensions (referred to as Level 1), formally specified as:

$Perceive \sqsubseteq Empathy\_dimension$

$Act \sqsubseteq Empathy\_dimension$

$Manifest \sqsubseteq Empathy\_dimension$

$Theory\_of\_mind \sqsubseteq Empathy\_dimension$

$Ratification \sqsubseteq Empathy\_dimension$

$Interpersonal \sqsubseteq Empathy\_dimension$

The main dimensions are comprised of sets of sub-dimensions (referred to as Level 2).

$Perceive\_Behavior \sqsubseteq Perceive$

$Perceive\_Context \sqsubseteq Perceive$

$Perceive\_Emotion \sqsubseteq Perceive$

$Perceive\_Interaction\_property \sqsubseteq Perceive$

$Perceive\_Message \sqsubseteq Perceive$

$Perceive\_Mood \sqsubseteq Perceive$

$Act\_Behavior \sqsubseteq Act$

$Act\_Context \sqsubseteq Act$

$Act\_Emotion \sqsubseteq Act$

$Act\_Interaction\_property \sqsubseteq Act$

$Act\_Message \sqsubseteq Act$
$Act\_Mood \sqsubseteq Act$

$Manifest\_Agreements \sqsubseteq Manifest$
$Manifest\_Emotion \sqsubseteq Manifest$
$Manifest\_Errors \sqsubseteq Manifest$
$Manifest\_Goals \sqsubseteq Manifest$
$Manifest\_Mood \sqsubseteq Manifest$
$Manifest\_Needs \sqsubseteq Manifest$
$Manifest\_Personality \sqsubseteq Manifest$
$Manifest\_Rapport \sqsubseteq Manifest$
$Manifest\_Social\_bond \sqsubseteq Manifest$
$Manifest\_Trust \sqsubseteq Manifest$
$Manifest\_Values \sqsubseteq Manifest$

$Tom\_Agreements \sqsubseteq Theory\_of\_mind$
$Tom\_Emotion \sqsubseteq Theory\_of\_mind$
$Tom\_Errors \sqsubseteq Theory\_of\_mind$
$Tom\_Goals \sqsubseteq Theory\_of\_mind$
$Tom\_Mood \sqsubseteq Theory\_of\_mind$
$Tom\_Needs \sqsubseteq Theory\_of\_mind$
$Tom\_Personality \sqsubseteq Theory\_of\_mind$
$Tom\_Rapport \sqsubseteq Theory\_of\_mind$
$Tom\_Social\_bond \sqsubseteq Theory\_of\_mind$
$Tom\_Trust \sqsubseteq Theory\_of\_mind$
$Tom\_Values \sqsubseteq Theory\_of\_mind$

$Ratification\_Acceptance \sqsubseteq Ratification$
$Ratification\_Agreements \sqsubseteq Ratification$
$Ratification\_Equality \sqsubseteq Ratification$
$Ratification\_Mutuality \sqsubseteq Ratification$

$Interpersonal\_Interconnection \sqsubseteq Interpersonal$
$Interpersonal\_Social\_Bonding \sqsubseteq Interpersonal$
$Interpersonal\_Tolerance \sqsubseteq Interpersonal$
$Interpersonal\_Trust_{b}uilding \sqsubseteq Interpersonal$

These sub-dimensions are further defined by their sub-dimensions (referred to as Level 3). For instance, examples of sub-dimensions of $Perceive\_Message \sqsubseteq Perceive$ are:

$Perceive\_Body\_language \sqsubseteq Perceive\_Message$
$Perceive\_Facial\_Expressions \sqsubseteq Perceive\_Message$
$Perceive\_Presence \sqsubseteq Perceive\_Message$
$Perceive\_Speech \sqsubseteq Perceive\_Message$
$Perceive\_Text \sqsubseteq Perceive\_Message$

The ontology's hierarchical structure provides explicit definitions of concepts at lower levels, enhancing assessments by enabling more objectively understood concepts. This promotes interoperability and conceptual alignment between systems that require a shared understanding of empathy concepts.

Furthermore, the hierarchical structure aids in human-system conceptual alignment. When gathering a human's perceptions about empathy concepts, it is crucial that the human's understanding aligns with the intended meaning of the ontology. Therefore, the ontology must offer precise and unambiguous means for comprehending observations, benefiting both systems and humans. Through the definition of concepts in this hierarchical structure, the ontology facilitates both system-to-system and human-to-system conceptual alignment.

Given a set of subjective perceptions from different sources, we assume that the dimensions of computational empathy regard imprecise perceptions, which define intervals between an upper bound and a lower bound. In order to capture imprecise perceptions, for each dimension and sub-dimension, we assign properties for imprecise uncertainty intervals. In this way, the ontology allows us to represent and reason about *empathic performance* in terms of imprecise measures.

5.3. Performance measures and imprecise uncertainty intervals

Imprecise perceptions are captured in the model in terms of imprecise uncertainty intervals (Dubois & Prade, 1998), which values can be obtained from sets of subjective perceptions of an interactive agent, consisting of an upper bound (most optimistic perception) and a lower bound (most skeptical perception). Hence, for each empathy dimension, given a scale $S \in [0, 1]$, there is an imprecise uncertainty interval I between an upper bound $t \in [0, 1]$ and a lower bound $b \in [0, 1]$, where $t \geq b$. This gives an imprecise measure of the performance in terms of the interval I, where the width of I represents the agreement/accuracy of the measures. An interval with upper bound and lower bound closer to 1.0 is understood as a higher performance measure, e.g., an accurately perceived emotion. This representation allows us to analyze sub-dimensions in terms of *performance* and *accuracy*. Furthermore, through the ontology's taxonomy, intervals of sub-dimensions can be aggregated to higher level dimensions to reason about performance in the 6 main dimensions.

We implement the intervals by defining a set of classes for managing weighted concepts:

- Class $Uncertainty\_interval$
  Properties: *hasLowerBound*, *hasUpperBound*
  Sub class of: *EmPAT:Concept*
- Data Property *hasUpperBound*
  Domain: $Uncertainty\_interval$
  Range: xsd:decimal
- Data Property: *hasLowerBound*
  Domain: $Uncertainty\_interval$
  Range: xsd:decimal
- Object Property: *hasInterval*
  Domain: $Empathy\_dimension$
  Range: $Uncertainty\_interval$

With the ontology's taxonomy as a platform, it is essential to capture interdependencies among concepts (semantic relations) to facilitate advanced reasoning on empathy observations. This necessity arises from the inherent challenge of being unable to directly observe an agent's internal empathic processes. As an example, the model must capture relations between "observable concepts", such as Act, and "internal concepts", such as $Manifest$ and $Theory\_of\_mind$.

5.4. Semantic interdependencies between empathy dimensions

By considering theories about empathy, e.g. social cognition (Szanto & Krueger, 2019), the Perception and Action Model (PAM) (Preston, 2007) and the Interpersonal Adaptation Theory (IAT) (Burgoon et al., 1995), we can define semantic relations among empathy concepts.

According to the Perception and Action Model (PAM), there is a chain of dependencies between perception, affective and cognitive processes, and action (see a summary in Section 2). Let us specify an abstraction of PAM in terms of the proposed 6 main dimensions of the current work. Perceptions lead to *affection* of an agent's mental/affective state ($Manifest$) and is *reflected* in its understanding of the other/cognitive state ($Theory\_of\_mind$). An agent's acting (Act) is in turn *influenced* by the affective processes and *tailored* by cognitive processes ($Manifest$ and $Theory\_of\_mind$). In order to capture

interrelations between mental and social processes, we further look at principles from social cognition (Szanto & Krueger, 2019), where empathy has been considered an important construct for interconnected and shared emotion (see a summary in Section 2). Social cognition, in terms of the proposed model, is an abstraction where $Manifest$ *enables* $Interpersonal$ (e.g., $Interconnected\_Emotions$), and $Theory\_of\_mind$ is *required* for $Ratification$, where an agent's Act is *forming* $Ratification$ (e.g., reaching social $Acceptance$ and $Agreements$). An agent's behavior has further relevance for adapting the $Interpersonal$ relation, explained by principles from the Interpersonal Adaptation Theory (IAT) (Burgoon et al., 1995). In terms of the proposed model, we specify an abstraction such that Act is *adapting* $Interpersonal$.

Before formalizing these relations, let us begin by specifying some relevant object (semantic) properties:

- Object Property: *isAffecting* - Perceive is affecting manifest.
  Domain: $Perceive$
  Range: $Manifest$
- Object Property: *isReflectedIn* - Perceive is reflected in theory of mind.
  Domain: $Perceive$
  Range: $Theory\_of\_mind$
- Object Property: *isEnabling* - Manifest is enabling interpersonal capability.
  Domain: $Manifest$
  Range: $Interpersonal$
- Object Property: *isRequiredFor* - Theory of mind is required for ratification.
  Domain: $Theory\_of\_mind$
  Range: $Ratification$
- Object Property: *isInfluencing* - Manifest is influencing an agent's acts.
  Domain: $Manifest$
  Range: Act
- Object Property: *isTailoring* - Theory of mind is tailoring an agent's acts.
  Domain: $Theory\_of\_mind$
  Range: Act
- Object Property: *isForming* - Act is forming ratification.
  Domain: Act
  Range: $Ratification$
- Object Property: *isAdapting* - Act is adapting Interpersonal.
  Domain: Act
  Range: $Interpersonal$

The chain of dependencies can be formalized as follows.

(1) Perceive *isAffecting* Manifest. An agent's perception *is affecting* the agent's manifest. For instance, a perceived emotion will affect the agent's emotion model. Hence, we say that a concept c that is perceived ($c \sqsubseteq Perceive$) will also have a representation in manifest ($c \sqsubseteq Manifest$). This is specified by the relation *isAffecting*, formally defined as:

$$\exists isAffecting.Manifest \sqsubseteq Perceive \sqcap Manifest \tag{1a}$$

By considering this dependency between Perceive and Manifest, we say that an agent has an *Empathic Manifest* if $isAffecting.Manifest$ holds for all perceptions:

$$EmpathicManifest \sqsubseteq Perceive \sqcap Manifest \sqcap \forall isAffecting.Manifest \tag{1b}$$

(2) Perceive *isReflectedIn* Theory_of_mind. An agent's perception *is reflected in* the agent's theory of mind. For instance, a perceived emotion will be reflected on by the agent. Hence, we say that a concept c that is perceived ($c \sqsubseteq Perceive$) will also have a representation in theory of mind ($c \sqsubseteq Theory\_of\_mind$). This is specified by the relation *isReflectedIn*, formally defined as:

$$\exists isReflectedIn.Theory\_of\_mind \sqsubseteq Perceive \sqcap Theory\_of\_mind \tag{2a}$$

By considering this dependency between Perceive and Theory of mind, we say that an agent has an *Empathic ToM* if $isReflectedIn.Theory\_of\_mind$ holds for all perceptions:

$$EmpathicToM \sqsubseteq Perceive \sqcap Theory\_of\_mind \sqcap \forall isReflectedIn.Theory\_of\_mind \tag{2b}$$

(3) Manifest *isEnabling* Interpersonal. An agent's manifest *is enabling* the agent's capability for interpersonal relation, where, e.g., an agent's manifested emotions may affect the other. Hence, we say that if an agent has manifested a concept c ($c \sqsubseteq Manifest$) such that $EmpathicManifest$ holds, it will also have a representation in Interpersonal ($c \sqsubseteq Interpersonal$), e.g., in the sub-dimension $Interconnected\_Emotion$. This is specified by the relation *isEnabling*, formally defined as:

$$\exists isEnabling.Interpersonal \sqsubseteq EmpathicManifest \sqcap Interpersonal \tag{3a}$$

By considering this dependency between Manifest and Interpersonal, we say that an agent has an *Empathic Interpersonal* if $isEnabling.Interpersonal$ holds for all $EmpathicManifest$:

$$EmpathicInterpersonal \sqsubseteq EmpathicManifest \sqcap Interpersonal \sqcap \forall isAffecting.Interpersonal \tag{3b}$$

(4) Theory_of_mind *isRequiredFor* Ratification. An agent's theory of mind *is required for* the agent's capability for ratification, where, e.g., an agent's theory of trust is central for reaching agreements. Hence, we say that if an agent has theory of mind of a concept c ($c \sqsubseteq Theory\_of\_mind$) such that $EmpathicToM$ holds, it will also have a representation in Ratification ($c \sqsubseteq Ratification$), e.g., in the sub-dimensions of $Trust$ and $Agreement$, respectively. This is specified by the relation *isRequiredFor*, formally defined as:

$$\exists isRequiredFor.Ratification \sqsubseteq Theory\_of\_mind \sqcap Ratification \tag{4a}$$

By considering this dependency between Theory of mind and Ratification, we say that an agent has an *Empathic Ratification* if the relation $isRequiredFor.Interpersonal$ holds for all $EmpathicToM$:

$$EmpathicRatification \sqsubseteq EmpathicToM \sqcap Ratification \sqcap \forall isRequiredFor.Ratification \tag{4b}$$

(5) Manifest *isInfluencing* Act. An agent's manifest *is influencing* the agent's act. For instance, a manifested emotion will influence the agent's emotional acting. Hence, we say that a concept c that is manifested ($c \sqsubseteq Manifest$) will also have a representation in act ($c \sqsubseteq Act$). This is specified by the relation *isInfluencing*, formally defined as:

$$\exists isAffecting.Act \sqsubseteq Manifest \sqcap Act \tag{5a}$$

By considering this dependency between Manifest and Act, we say that an agent has an *Empathic Act* if the relation $isInfluencing.Act$ holds for all $EmpathicManifest$:

$$EmpathicAct \sqsubseteq EmpathicManifest \sqcap Act \sqcap \forall isInfluencing.Act \tag{5b}$$

(6) Theory_of_mind *isTailoring* Act. An agent's Theory of mind *is tailoring* the agent's act. For instance, a theory of another agent's emotion will tailor the agent's emotional acting. Hence, we say that a concept c that is in theory of mind ($c \sqsubseteq Theory\_of\_mind$) will also have a representation in act ($c \sqsubseteq Act$). This is specified by the relation *isTailoring*, formally defined as:

$$\exists isTailoring.Act \sqsubseteq EmpathicToM \sqcap Act \tag{6a}$$

By considering this dependency between Theory of mind and Act, we say that an agent has an *Empathic Act* if the relation $isTailoring.Act$ holds for all $EmpathicToM$:

$$EmpathicAct \sqsubseteq EmpathicToM \sqcap Act \sqcap \forall isTailoring.Act \tag{6b}$$

(7) Act *isForming* Ratification. An agent's Act *is forming* the agent's ratification with other agents. For instance, the agent's emotional acting will form emotional acceptance. Hence, we say that a concept c that is in an empathic act ($c \sqsubseteq EmpathicAct$) will also have a representation in ratification ($c \sqsubseteq Ratification$). This is specified by the relation *isForming*, formally defined as:

$$\exists isForming.Ratification \sqsubseteq EmpathicAct \sqcap Ratification \tag{7a}$$

By considering this dependency between EmpathicAct and Ratification, we say that an agent has an *Empathic Ratification* if the relation $isForming.Ratification$ holds for all $EmpathicAct$:

$$EmpathicRatification \sqsubseteq EmpathicAct \sqcap Ratification \sqcap \forall isForming.Ratification \tag{7b}$$

(7) Act *isAdapting* Interpersonal. An agent's Act *is adapting* the agent's interpersonal relation with other agents. For instance, the agent's emotional acting will adapt the emotional connection between agents. Hence, we say that a concept c that is in an empathic act ($c \sqsubseteq EmpathicAct$) will also have a representation in interpersonal ($c \sqsubseteq Interpersonal$). This is specified by the relation *isAdapting*, formally defined as:

$$\exists isAdapting.Interpersonal \sqsubseteq EmpathicAct \sqcap Interpersonal \tag{8a}$$

By considering this dependency between EmpathicAct and Interpersonal, we say that an agent has an *Empathic Interpersonal* if the relation $isAdapting.Interpersonal$ holds for all $EmpathicAct$:

$$EmpathicInterpersonal \sqsubseteq EmpathicAct \sqcap Interpersonal \sqcap \forall isAdapting.Interpersonal \tag{8b}$$

The above specifications allow a semantic understanding of the defined concepts of computational empathy. The semantic interdependencies are important for reasoning about concepts that may not be directly observed nor inferred through the taxonomy. The semantic interdependencies go beyond the taxonomy to reason about an agent's perceived empathic capability. Such perceptions can be transformed into DL queries, provided as input to the ontology to give further explanations of the perceptions.

5.4.1. Examples of reasoning queries

These semantic interdependencies enable the ontology to be applied for reasoning about abstract empathy concepts. A collection of low level perceptions of the interaction allows to make conclusions about how an agent's empathic capabilities are perceived. For instance, given the perception: "an agent seems to be sensing and acting by considering emotions", the following DL query can be created: "Perceive_Emotion ⊓ Act_Emotion ⊓ isAffecting.Act". Given this query, we can, by considering the above chain of semantic interdependencies, infer "EmpathicManifest ⊓ EmpathicToM", estimating that the agent was perceived to be mentally affected by, and to reason about, emotions. Given this assertion, we can advance with further queries, such as: "EmpathicManifest ⊓ EmpathicToM" to infer "EmpathicAct ⊓ (isInfluencing.Act ⊔ isTailoring.Act)", estimating that an agent was perceived to act empathically. By aggregating the associated uncertainty intervals of asserted concepts, explicit performance measures can be estimated. For instance, measured uncertainty intervals of observed acts, e.g., Act_Emotion and Act_Context, with upper bounds ∈ [0,1] and lower bounds ∈ [0,1], are aggregated to corresponding uncertainty intervals of higher level concepts, e.g., Act, estimating the performance of abstract computational empathy concepts. The interdependencies further allow to infer aggregated uncertainty intervals of semantically related concepts, e.g., interpersonal and ratification.

In this section, we have formally defined a model for computational empathy. A taxonomy is informed by a grounded theory study on data collected in a literature review of empathy. The interdependencies between concepts, for reasoning about ramification effects, are informed by prior empathy theories (e.g., PAM (Preston, 2007) and social cognition (Szanto & Krueger, 2019)). In the following section, we present the process and results of a user study where we apply the model as a tool for reasoning about perceptions of empathy in two state of the art chatbots: Replika and Wysa.

6. Example application: Assessing perceptions of chatbot empathy

Our primary goal is to establish a broad, versatile definition of computational empathy that can be tailored to specific applications as needed. In the following example, applying the proposed model of computational empathy as a tool for assessing empathy in chatbots, we illustrate how we can uncover trends in changing perceptions of computational empathy over time. By correlating these trends with low-level concepts within the ontology, detailed insights into observed changes can be provided.

In order to enable data collection for assessing empathy in chatbots, an assessment protocol (in the form of a questionnaire) is developed, consisting of Likert scale (1–5) questions designed to collect information corresponding to the dimensions of computational empathy. A user-study was conducted using the tool, in which the participants interacted with two state-of-the-art chatbots in the area of health and well-being (Replika and Wysa). The user-study collected the users' perceptions of the chatbots' interaction capabilities in terms of the lowest level dimensions (Level 3; the most trivial concepts regarding understanding and measuring) of the empathy ontology. The answers were normalized ([0,1]) and clustered on different levels in the ontology to derive qualitative *empathy measures*. The measures were analyzed on an aggregated level, in terms of the 6 main dimensions, as well as on a detailed level, considering each sub-dimension's contribution.

Recall that the ontology is structured according to a taxonomy where general, more abstract, concepts are at a high level and more specialized, less abstract, concepts are at a low level. By designing the questionnaire questions based on low level concepts in the ontology, the questionnaire acts as an interface between the ontology and the users' perceptions. The 6 high level dimensions of the ontology (Level 1), such as "Theory of Mind" and "Perceive", are abstract concepts, difficult to intuitively understand for a participant. An answer to such a high level question would be based on the participant's subjective meaning of the concept, which may differ from what is intended in the questionnaire. A level down the taxonomy (Level 2), more understandable, intermediately abstract, concepts are reached, such as "Emotion" and "Message", which still may be subjectively understood. By continuing further down the taxonomy (Level 3), we reach concepts such as "Response rate", "Facial expressions" and "Body language", which are more explicit and more intuitively understood by a participant. This makes it easier to map the intended meaning of the concepts to questions in the questionnaire. As a result, the answers are less subjective, making them more trivial to measure and analyze. This method of subdividing concepts to reach more trivial concepts can be applied in any number of levels, until a sufficient formality is reached. This study stopped at Level 3, to balance formality and detail.

6.1. User study design

The goal of the study is to illustrate the capability of the model to describe users' perceptions of computational empathy on different chatbots (which have different interaction capabilities). Preferably,

when conducting such a study, we would have a configurable chatbot which could enable/disable different empathic features (e.g., actuators and sensors) to create different experimental conditions. However, let us mention that there are no models in the state of the art to configure an empathic chatbot. Alternatively, by designing the study using two different chatbots, Wysa and Replika (see Fig. 4),[4,5] with a diverse set of features, we create different experimental conditions, whose empathetic features can be traced back to design decisions of the chatbots (see Table 7). The chatbot features are selected based on two classes: *Actuators* (social actions) and *Sensors* (social perceptions). These are aspects that we can objectively identify in the chatbots, which otherwise can be understood as "blackbox" systems. Given these features, we assess *performance measures* in terms of the dimensions of computational empathy. We can understand the chatbot features as objective truths about the chatbots. For instance, Replika has an actuator in terms of facial expressions, Wysa does not. Hence, some performance measures for sub-dimensions, such as $Act\_Facial\_expressions$, are conceptually and functionally linked to this feature. This allow us to explain the participants' answers more objectively as well as tracing aspects of empathy to explicit design features of the chatbots. We distinguish our experimental conditions (the two chatbots) by conducting human–chatbot interaction "in the wild", providing a realistic setting for assessing users' perceptions of chatbot empathy.

Given the relations between chatbot features and empathy dimensions (Table 7), by considering actuators and sensors for Wysa in comparison to Replika, we can specify a set of hypotheses for the chatbots' performances, and by considering the interdependencies defined in the semantic model, we can specify an argument chain for each hypothesis:

- Hypothesis 1 (H1): Measures of ***Act*** on day 3 and 7 will be lower for Wysa than Replika
  - Argument for H1: Wysa lacks actuator A1, A3, A4, A5, which limit Act
  - Argument for H1: Wysa lacks sensor S2, S3, S4, which limit Perceive
  - Argument for H1: Perceive is affecting Manifest
  - Argument for H1: Perceive is reflected in Theory of mind
  - Argument for H1: Manifest is influencing Act
  - Argument for H1: Theory of mind is tailoring Act
- Hypothesis 2 (H2): Measures of ***Manifest*** on day 3 and 7 will be lower for Wysa than Replika
  - Argument for H2: Wysa lacks sensor S2, S3, S4, which limit Perceive
  - Argument for H2: Perceive is affecting Manifest
- Hypothesis 3 (H3): Measures of ***Perceive*** on day 3 and 7 will be lower for Wysa than Replika
  - Argument for H3: Wysa lacks sensor S2, S3, S4, which limit Perceive
- Hypothesis 4 (H4): Measures of ***Theory of mind*** on day 3 and 7 will be lower for Wysa than Replika
  - Argument for H4: Wysa lacks sensor S2, S3, S4, which limit Perceive
  - Argument for H4: Perceive is reflected in Theory of mind
- Hypothesis 5 (H5): Measures of ***Interpersonal*** between day 3 and 7 will be more increased for Wysa than Replika
  - Argument for H5: Wysa lacks actuator A1, A3, A4, A5, which limit Act
  - Argument for H5: Wysa lacks sensor S2, S3, S4, which limit Perceive
  - Argument for H5: Perceive is affecting Manifest
  - Argument for H5: Manifest is enabling Interpersonal
  - Argument for H5: Act is adapting Interpersonal
- Hypothesis 6 (H6): Measures of ***Ratification*** on day 3 and 7 will be lower for Wysa than Replika
  - Argument for H6: Wysa lacks actuator A1, A3, A4, A5, which limit Act
  - Argument for H6: Wysa lacks sensor S2, S3, S4, which limit Perceive
  - Argument for H6: Perceive is reflected in Theory of mind
  - Argument for H6: Theory of mind is required for Ratification
  - Argument for H6: Act is forming Ratification

Selection of participants. The selection aimed for university students, below age of 45 (young adult) and with minimal prior chatbot experience. In total 13 participants were selected. The participant group were mostly current or prior university students in a variety of study areas (digital design, philosophy, cognitive science, computing science, law, healthcare, and engineering). Given the similarity in nature between social media chats and chatbot interactions, we collected the participants' social media chat experience in addition to prior chatbot experience. This was done in order to estimate a sufficient capability among participants of handling a chat interface. The participants generally described their prior experience with chatbot interactions as low to medium and social media interactions as medium to high (see Table 6).

Procedure. We clustered the 13 participants into two groups, one to interact with Replika and the other to interact with Wysa. The participants were asked to interact with the chatbot for about 10 min, once per day, for a week. On two occasions (on day three, N=13, and day seven, N=12), the participants were asked to answer an online questionnaire about the interaction experience. The two rounds of the questionnaire were done in order to collect a measure of how a long-term interaction affects the empathy scores. We clarified that the participant decides what to share with the chatbot, and that we will not ask to see the chat log. Making the interaction private was important in order to promote the participants to comfortably decide topics with the chatbot.

Assessment protocol. The protocol has Likert scale questions in a 1–5 range (see Tables 8 and 9), covering the lower level dimensions of empathy. The 1–5 alternatives on each question represent levels of performance regarding a particular empathic capability, where 1 represents low performance and 5 represents high performance.

Limitations. This user study aims to illustrate trends in changing perceptions of computational empathy over time, presenting a potential application of the ontology. Considering the constraints of our sample size, our intent is to provide a starting point for understanding differences in empathy dimensions. For these particular chatbots, the presented trends should be viewed as preliminary and subject to further investigation.

6.2. Initial analysis (average)

The 1–5 data range is normalized to values between 0 and 1. In order to get an overview of the data to find overall patterns, an initial analysis was done by calculating average values (arithmetic mean) for each question (sub-dimension), separately for each chatbot. The calculated averages for each sub-dimension was then further aggregated

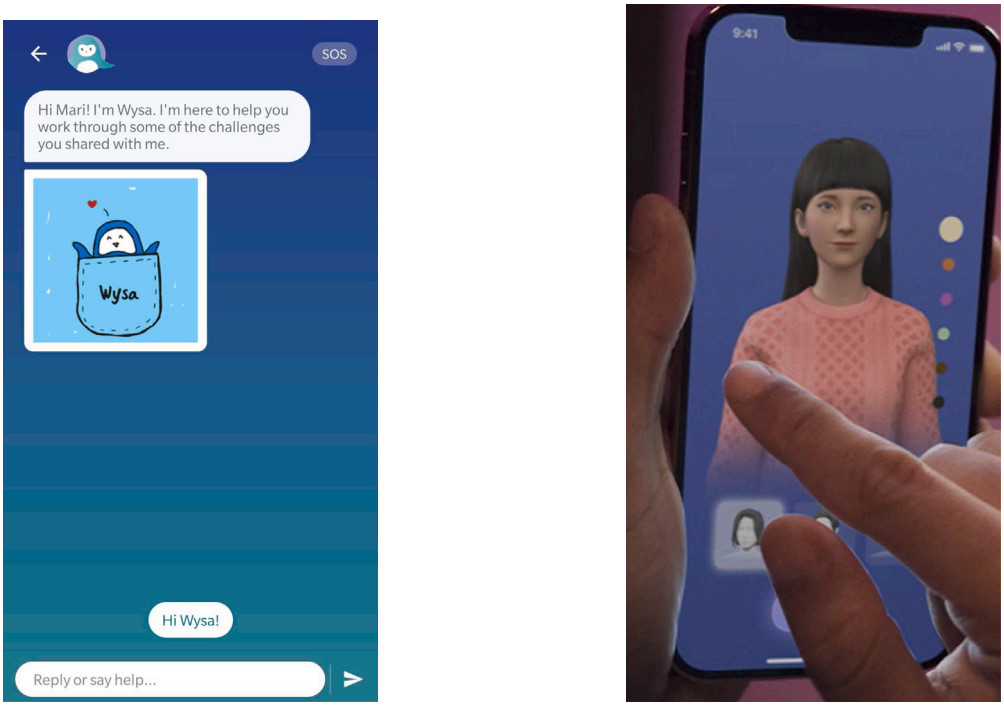
[4] Press kit Wysa: https://www.wysa.io/media.
[5] Press kit Replika: https://replika.com/about/press.

Table 6
Participants of the user study.

Participant	Age	Gender	Exp Social Media (1-5)	Exp Chatbots (1-5)	Study Area	Chatbot
W21	35	M	2	1	Computing Science	Wysa
W25	25	F	5	1	Cognitive Science	Wysa
W27	33	F	3	3	Cognitive Science	Wysa
W29	29	M	4	3	Cognitive Science	Wysa
W30	26	M	4	2	Computing Science	Wysa
W36	29	M	5	1	Philosophy	Wysa
R22	41	M	5	2	Computing Science	Replika
R23	29	F	2	1	Digital Design	Replika
R24	32	F	5	1	Cognitive Science	Replika
R31	36	M	4	1	Law	Replika
R33	28	M	3	3	Healthcare	Replika
R35	33	M	3	1	Psychology	Replika
R37	60+	M	3	1	Engineering	Replika

Table 7
Experimental conditions in terms of chatbot features. Features are related to empathy dimensions.

ID	Feature	Wysa	Replika	Related Empathy Dimensions
A1	Actuator: Human appearance		X	Act
A2	Actuator: Text messages	X	X	Act
A3	Actuator: Facial expressions		X	Act
A4	Actuator: Body language		X	Act
A5	Actuator: Affective/emotional expressions		X	Act, Manifest
S1	Sensor: Natural language text messages	X	X	Perceive, ToM
S2	Sensor: Emojis; User writes, e.g., :) :/ :(		X	Perceive, ToM
S3	Sensor: Roleplay; User writes an action, e.g., *Looks at you*		X	Perceive, ToM
S4	Sensor: User message count; Unlocking traits on certain message counts		X	Perceive, Interpersonal



(a) Wysa (b) Replika

Fig. 4. Chatbots constitute different experimental conditions (pictures retrieved from press kits; footnote 5 and 6).

to calculate an average for each of the 6 main dimensions. Finally, a total average value for each chatbot is calculated. These averages give a rough estimate about the common values in the data set and an approximation of the performance measures (empathy scores) for each chatbot. We now present the initial analysis of average performance measures w.r.t. the dimensions on Day 3 and Day 7. In each stage of the analysis, we trace consistencies between the chatbot features (Table 7) and the participants' perceptions of the chatbot interaction.

6.2.1. *Average dimension measures on day 3*

On day three (see Fig. 5(a)), we can observe overall higher scores for Replika total average 0.52 than for Wysa total average 0.36. Looking

Table 8
Part I: Questionnaire questions. Questions are mapped to sub-dimensions of the ontology.

Empathy dimension	Question (1-5 Likert scale)
Perceive Mood	1. On what level did the Chatbot seem to perceive your current mood?
Perceive Emotion	2. On what level did the Chatbot seem to perceive emotions you expressed?
Perceive Behavior	3. On what level did the Chatbot seem to perceive your behavior?
Perceive Message	4. On what level did the Chatbot seem to perceive relevant information?
Perceive Demography	5. On what level did the Chatbot seem to perceive your age, gender and culture?
Perceive Activity	6. On what level did the Chatbot seem to perceive your purpose of the dialogue?
Perceive Response Rate	7. On what level did the Chatbot seem to perceive your response rate?
Perceive Time	8. On what level did the Chatbot seem to perceive time between sessions?
Perceive Duration	9. On what level did the Chatbot seem to perceive duration of interaction?
Perceive Response Content Level	10. On what level did the Chatbot seem to perceive the amount of text of your response?
ToM Agreements	11. On what level did the Chatbot seem to understand established agreements?
ToM Personality	12. On what level did the Chatbot seem to understand your personality?
ToM Emotion	13. On what level did the Chatbot seem to understand your emotions?
ToM Rapport	14. On what level did the Chatbot seem to understand your rapport?
ToM Trust	15. On what level did the Chatbot seem to understand your trust?
ToM Errors	16. On what level did the Chatbot seem to understand your errors?
ToM Deception	17. On what level did the Chatbot seem to understand deception?
ToM Social Bond	18. On what level did the Chatbot seem to understand what type of relation you had?
ToM Values	24. On what level did the Chatbot seem to understand your values?
ToM Goals	25. On what level did the Chatbot seem to understand your goals?
ToM Needs	26. On what level did the Chatbot seem to understand your needs?
Act Mood	27. On what level did the Chatbot express appropriate mood?
Act Emotion	28. On what level did the Chatbot express appropriate emotions?
Act Behavior	29. On what level did the Chatbot express appropriate behavior?
Act Message	30. On what level did the Chatbot express relevant information?
Act Demography	31. On what level did the Chatbot act appropriately according to your demography?
Act Activity	32. On what level did the Chatbot act in relation to your purpose of the dialogue?
Act Response Rate	33. On what level did the Chatbot act in relation to your response rate?
Act Time	34. On what level did the Chatbot act in relation to time between sessions?
Act Duration	35. On what level did the Chatbot act in relation to the duration of interaction?
Act Response Content	36. On what level did the Chatbot act in relation to the size of your responses?
Manifest Emotion	37. On what level was the Chatbot's emotions affected by your emotions?
Manifest Personality	38. On what level was the Chatbot's personality appropriately affected?
Manifest Trust	40. On what level did the Chatbot appear respectful?
Manifest Trust	41. On what level did the Chatbot appear responsible?
Manifest Trust	42. On what level did the Chatbot appear fair?
Manifest Trust	43. On what level did the Chatbot appear honest?
Manifest Behavior	44. On what level was the Chatbot's behavior appropriately affected?
Manifest Values	45. On what level did the Chatbot show signs of behaving in line with certain values?
Manifest Goals	46. On what level did the Chatbot show signs of behaving in line with certain goals?
Manifest Needs	47. On what level did the Chatbot show signs of behaving in line with certain needs?

at averages for the 6 main dimensions, we can observe the following average differences (comparing Replika/Wysa): Perceive (0.44/0.35), Act (0.52/0.23), ToM (0.47/0.32), Manifest (0.60/0.47) Ratification (0.51/0.43), Interpersonal (0.57/0.37), showing that each of the 6 main dimensions were higher for Replika. This rough estimate is analyzed in detail in Section 6.4.

6.2.2. Average dimension measures on day 7

On day seven (see Fig. 5(b)), we can observe overall higher scores for Replika total average 0.49 than for Wysa total average 0.42. Looking at averages for the 6 main dimensions, we can observe the following average differences (comparing Replika/Wysa): Perceive (0.42/0.36), Act (0.42/0.46), ToM (0.42/0.33), Manifest (0.61/0.54), Ratification (0.52/0.41), Interpersonal (0.54/0.42), showing that most of the 6 main dimensions were still higher for Replika, except for Act that on day 7 is higher for Wysa. Act notably increased for Wysa, from 0.23 (day 3) to 0.46 (day 7), and Act notably decreased for Replika, from 0.52 (day 3) to 0.42 (day 7). This approximation of change shows that the long-term interaction had an impact on the measures. These rough estimates of change between day 3 and day 7 can be explained by analyzing the individual sub-dimensions of e.g., Act between Day 3 and Day 7, finding dominant sub-dimensions that gave rise to the change. The sub-dimensions suggest the participants' initial perceptions (expectations) of the various chatbots (measures on Day 3) which were either met or rejected with further interaction (measures on Day 7). These causes of change are analyzed in detail in Section 6.4.

6.3. Analysis of possibilistic intervals

An analysis is done looking at intervals between maximum and minimum values of each dimension. This is done through an imprecise uncertainty measure considering intervals of possibilistic distributions (Dubois & Prade, 1998). Possibilistic distributions are a class of fuzzy sets (Zadeh, 1978), which can be calculated for each dimension to capture qualitative data, making it a practical method to deal with uncertain empathy measures. In order to preprocess the collected data from the questionnaire to create possibilistic distributions, we collected, for each dimension, the most optimistic (highest) answers and calculated the average. Similarly, for each dimension, we collected the most skeptical (lowest) answers and calculated the average, in this way defining an interval for each dimension. Through this approach, we can understand and visualize the data in a more nuanced way, finding a skeptical lowest point and an optimistic highest point for each observation.

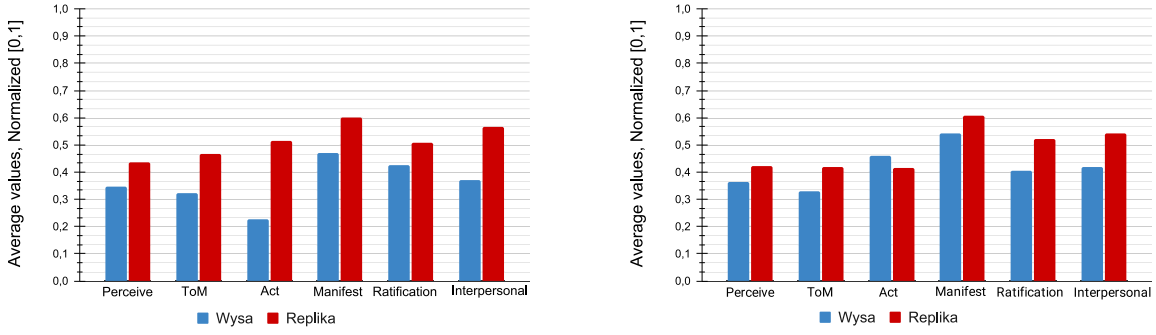
Looking at the intervals for Replika and Wysa (see Fig. 6), we can see that some intervals are more compact in terms of minimum and maximum (a higher agreement/accuracy of the participants' perceptions). We can further observe that the intervals generally became more compact between Day 3 and Day 7, showing that the long-term interaction streamlined the users' conceptions. Hence, the data on Day 7 can be regarded as less subjective.

6.3.1. Average dimension intervals on day 3

On day three (see Fig. 6(a)), we can observe a generally more diverse perspective from the participants, with less compact intervals,

Table 9
Part II: Questionnaire questions. Questions are mapped to sub-dimensions of the ontology.

Empathy dimension	Question (1-5 Likert scale)
Ratification Agreements	48. On what level did the Chatbot and you reach agreements?
Ratification Equality	49. On what level did the Chatbot and you have equality?
Ratification Mutuality	50. On what level did the Chatbot and you have mutuality?
Ratification Acceptance	51. On what level did the Chatbot and you accept each other?
Interpersonal Connected Mood	52. On what level did the Chatbot and you share mood?
Interpersonal Connected Emotion	53. On what level did the Chatbot and you share emotions?
Interpersonal Connected Behavior	54. On what level did the Chatbot and you share behavior?
Interpersonal Connected Message	55. On what level did the Chatbot and you share knowledge?
Interpersonal Trust Building	56. On what level did the Chatbot and you build trust?
Interpersonal Trust Building	57. The Chatbot honored commitments
Interpersonal Trust Building	58. The Chatbot communicated effectively and clear
Interpersonal Trust Building	59. The Chatbot made decisions in a careful way
Interpersonal Trust Building	60. The Chatbot behaved in a consistent way
Interpersonal Trust Building	61. The Chatbot is an active listener
Interpersonal Trust Building	62. The Chatbot provided respectful feedback
Interpersonal Trust Building	63. The Chatbot was open for your thoughts and feelings
Interpersonal Trust Building	64. The Chatbot showed feelings
Interpersonal Trust Building	65. The Chatbot was honest
Interpersonal Trust Building	66. The Chatbot was helpful
Interpersonal Trust Building	67. The Chatbot was kind in an genuine way
Interpersonal Trust Building	68. The Chatbot did not promote itself
Interpersonal Trust Building	69. The Chatbot recognized you as a person
Interpersonal Trust Building	70. The Chatbot behaved in line with certain values
Interpersonal Trust Building	71. The Chatbot admitted when it made a mistake
Interpersonal Trust Building	72. On what level did you accept the Chatbot despite its flaws?
Interpersonal Trust Building	73. On what level did the Chatbot and you build a relation?
Interpersonal Trust Building	74. On what level did the Chatbot and you build a professional relation, such as how it would be between doctor and patient?
Interpersonal Social Bond	75. On what level did the Chatbot and you build a family relation, such as how it would be between mother and daughter?
Interpersonal Social Bond	76. On what level did the Chatbot and you build a friend relation, such as how it would be between two close friends?
Interpersonal Social Bond	77. On what level did the Chatbot and you build a personal relation, such as how it would be between two co-workers?
Interpersonal Social Bond	78. On what level did the Chatbot and you build a social relation, such as how it would be between individuals at a party?



(a) Day 3 (b) Day 7

Fig. 5. Average empathy perceptions across the main empathy dimensions.

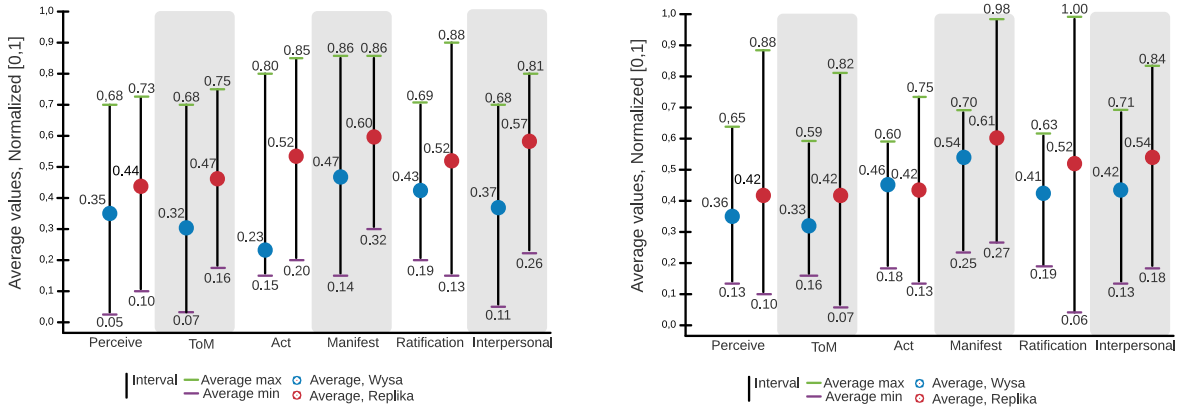
and with averages more toward the min/max endpoints. For instance, Wysa's Act (average 0.23) is close to the minimum (0.15) and far from the maximum (0.80). An interesting observation is that most of the dimensions had a higher maximum on day three, compared to day seven. This suggests an initially higher optimism from participants on the chatbots' interaction capabilities.

6.3.2. Average dimension intervals on day 7

On day seven (see Fig. 6(b)), we can observe more compact intervals in most of the dimensions, compared to day three. There are some exceptions, e.g., a higher spread on Replika's Ratification interval on day seven (max: 1.0, min: 0.06), compared to the same dimension on day three (max: 0.88, min: 0.13). The generally more compact intervals at day seven, compared to day three, suggest that the participants had a more common view on the interaction. This change can be explained in different levels by looking at different sets of dimensions. In the next subsection, we look at each sub-dimension for a detailed analysis and explanation.

6.4. Bottom-up analysis: Explaining change between day 3 and day 7

We now analyze the data from a bottom-up perspective. Generally, the bottom-up approach focuses its analysis on the lower concepts in the ontology descending the 6 main dimensions, to provide detailed explanations. Each sub-dimension in the ontology is mapped to questions in the questionnaire. Hence, by analyzing the collected responses through the ontology's taxonomy, we can get an aggregated understanding of the six main dimensions of empathy. Furthermore, by analyzing each sub-dimension, we can explain in detail why these changes, between day 3 and day 7, occurred. More dominant sub-dimensions which give rise to these changes can be recognized to ground the base for an explanation. Dominance of sub-dimensions can be calculated using different methods (e.g., in terms of difference, interval compactness, invariance, etc.). Here, we define dominance as the difference of maximum, and minimum, between day 3 and day 7. Let us look at the changed intervals to highlight dominant sub-dimensions in each chatbot study.



(a) Day 3 (b) Day 7

Fig. 6. Average empathy perceptions considering intervals (minimums and maximums) across the main empathy dimensions.

6.4.1. *Wysa sub-dimension intervals*

In the Wysa user study, by only considering the most skeptical (minimum) measures on each dimension (see Fig. 7(a)), we can observe that minimum values of *Act Mood*, *Act Purpose* and *Act Message* notably increased between day 3 and day 7 while minimum values of *Act Time* and *Act Duration* notably decreased. The sub-dimensions which made the most dominant change, w.r.t. their minimum measure was *Manifest Goals* which increased from 0.0 to 0.5 and *Ratification Acceptance* which decreased from 0.5 to 0.0.

Continuing our analysis of the Wysa user study, by only considering the most optimistic (maximum) measures on each dimension (see Fig. 7(b)) we can notice dominant changes between day 3 and day 7. Particularly notable changes are for sub-dimensions under *Act*, *Manifest* and *Theory of Mind*, where most of the maximum values increased, and sub-dimensions below *Interpersonal* where a majority of the maximum values decreased. For instance, notable increase of maximum values can be observed in *Act Mood*, *Act Emotions*, *Act Behavior*, *Act Content level*, *Manifest Content level*, *Manifest Emotions*, *Manifest Personality*, *Manifest Trust*, *Manifest Values*, *Manifest Needs* and *Manifest Agreements*. On the other hand, particular decreases can be observed in *Interpersonal Connected Mood*, *Interpersonal Connected Emotions*, *Interpersonal Connected Behavior*, *Perceive Purpose* and *Perceive Response Content level*. Hence, we can identify subsets of sub-dimensions that more dominantly changed between day 3 and day 7, explaining how the participants' perspectives changed w.r.t. the dimensions of computational empathy.

6.4.2. *Replika sub-dimension intervals*

In the Replika user study, by only considering the most skeptical (minimum) measures on each dimension (see Fig. 8(a)), we can observe that minimum values of *Manifest Emotions*, *Manifest Trust*, and *Theory of mind Goals* and *Perceive Mood* notably increased between day 3 and day 7 while minimum values of *Manifest Agreements*, *Theory of mind Values*, *Act Mood*, *Act Emotions* and *Act Purpose* notably decreased. The sub-dimensions which made the most dominant change w.r.t. their minimum measure were *Manifest Emotions*, increasing from 0.25 to 0.5, *Act Mood*, decreasing from 0.5 to 0.25, *Act Emotions*, decreasing from 0.5 to 0.25, and *Act Purpose*, decreasing from 0.5 to 0.0. Hence, we can make detailed estimates about which perceptions changed between day 3 and day 7. For instance, we can estimate that the participants had high expectations of Replika's purposeful acting at day 3, which was rejected through the long-term interaction (measured at day 7), a dominant factor of the greatly decreased measure of *Act*.

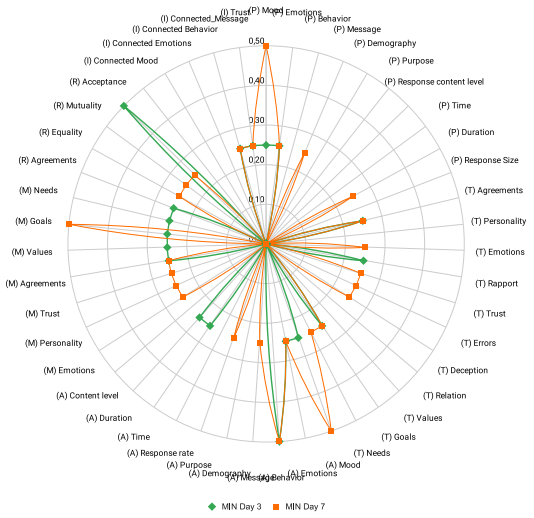
Continuing our analysis of the Replika user study, by only considering the most optimistic (maximum) measures on each dimension (see Fig. 8(b)), we can observe dominant changes between day 3 and day 7. Particularly notable changes are for sub-dimensions under *Manifest*, *Theory of Mind* and *Perceive*, where most of the maximum values increased, and sub-dimensions below *Act* and *Interpersonal* where a majority of the maximum values decreased. For instance, notable increase of maximum values can be observed in *Manifest Emotions*, *Manifest Values*, *Manifest Goals*, *Manifest Agreements*, *Perceive Mood*, *Perceive Emotions*, *Perceive Behavior*, *Perceive Response content level*, *Theory of mind Personality*, *Theory of mind Feelings*, *Theory of mind Trust*, *Theory of mind Goals*, *Theory of mind Needs* and *Interpersonal Trust building*, while notable decrease of maximum values can be observed in *Act Behavior*, *Act Message*, *Act Time*, *Interpersonal Connected mood*, *Perceive Duration* and *Theory of mind Relation*. By considering these observations, we can provide explanations about the chatbot's computational empathy. For instance, Replika was perceived to be emotionally affected by the interaction (given by *Manifest Emotions*) and had an increased understanding of the user's trust (given by *Theory of mind Trust*). Concurrently, Replika was perceived to have increased capability for trust building (given by *Interpersonal Trust building*). These observations give rise to aggregated as well as detailed explanations of the user's perceptions.

6.5. *Empathy performance measures related to chatbot features*

We now reconnect to the chatbot features presented in Table 7. Recall that two experimental conditions were defined as a configuration of features for each chatbot, Replika and Wysa. Each chatbot feature was further related to one or more dimensions of computational empathy. By considering these relations, six hypotheses (H1–H6) were specified. In the previous analyses, we processed the participants' perceptions of the chatbots' capabilities to identify dominant empathy dimensions. By mapping the perceived dimensions to chatbot features (according to Table 7) we can further analyze the data to make estimates about how these features had an impact on the perceptions.

By considering the hypotheses H1-H6 and the measured perceptions of empathy dimensions of Replika and Wysa, we can make the following observations:

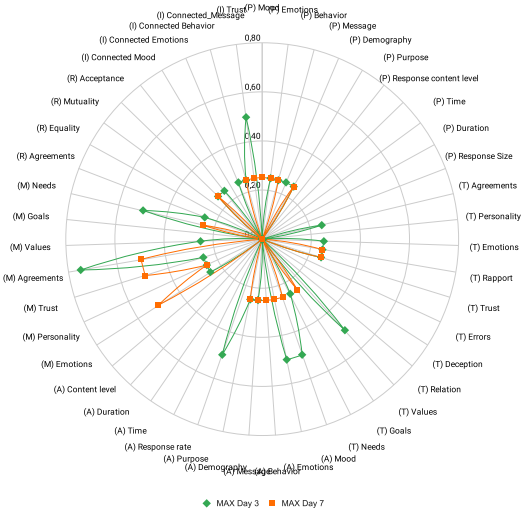
- H1 does not hold on Day 7. Wysa's Act was measured as higher than Replika's.
- H2 holds. Replika's Manifest was measured as higher than Wysa's.
- H3 holds. Replika's Perceive was measured as higher than Wysa's.
- H4 holds. Replika's Theory of mind was measured as higher than Wysa's.
- H5 holds. Replika's Interpersonal was measured as higher than Wysa's.
- H6 holds. Replika's Ratification was measured as higher than Wysa's.



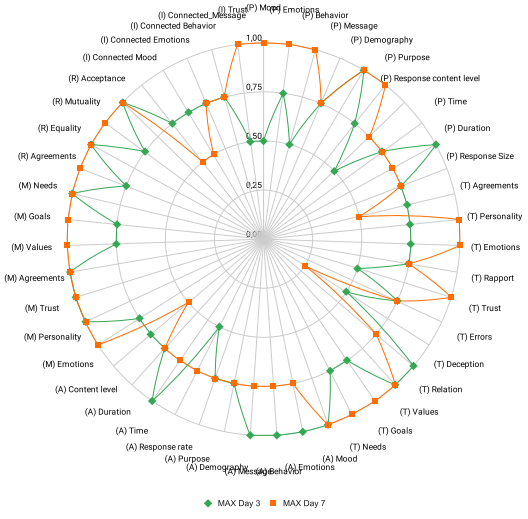
(a) Wysa: Minimums on Day 3 and Day 7

(b) Wysa: Maximums on Day 3 and Day 7

Fig. 7. Wysa: Changes in perceptions of empathy between Day 3 and Day 7.



(a) Replika: Minimums on Day 3 and Day 7



(b) Replika: Maximums on Day 3 and Day 7

Fig. 8. Replika: Changes in perceptions of empathy between Day 3 and Day 7.

By considering the specified chatbot features, we can observe that actuators and sensors, w.r.t. the defined dimensions, in most cases followed the hypotheses regarding the chatbots' perceived computational empathy. Such observations can provide directions for how a system can be designed for increasing this performance measure. An interesting observation is that Wysa's *Act* was perceived as higher than Replika's *Act* on day 7, although Wysa has less sensors and actuators than Replika. In our previous analyses, this is explained by specific sub-dimensions, where, e.g., *Act Emotions*, *Act Mood*, *Act Behavior* notably increased for Wysa, while sub-dimensions, such as *Act Time*, *Act Emotions*, *Act Mood*, *Act Behavior* notably decreased for Replika.

Let us look at how these observations can be further analyzed computationally. The ontology allows for alternative explanations of users' perceptions by considering higher-level empathy concepts indirectly derived from their responses. By identifying dominant dimensions in the measures, we can formulate queries to gain further insights into perceptions of computational empathy.

6.6. Reasoning queries on observed dominant dimensions

The semantic relations of the ontology can be utilized for reasoning about perceptions to generate theories about high level empathy concepts. Given a set of observations, we can create corresponding queries to the ontology. For instance, in the Replika user-study, we observed the dominant sub-dimensions *Perceive Emotions* and *Manifest Emotions*. These observations can be transformed into the following query: "$Manifest\_Emotion \sqcap Perceive\_Emotion \sqcap isAffecting.Manifest$". Then, given the defined formula (1b) $EmpathicManifest \sqsubseteq Perceive \sqcap Manifest \sqcap \forall isAffecting.Manifest$, we can infer "$EmpathicManifest$", asserting a theory stating that the chatbot was perceived to have a mental state affected by its perceptions. Furthermore, given the previous assertion, we can advance with additional queries, such as: "$EmpathicManifest \sqcap Manifest\_Emotion \sqcap Act$" to infer "$EmpathicAct \sqcap isInfluencing.Act \sqcap Act\_Emotion$", asserting theories about the chatbot's perceived capability of acting emphatically w.r.t. emotions. In this way, by considering the holistic semantic model of computational empathy, perceptions can be analyzed on a low level to reach conclusions about high level interdependent concepts, e.g., $EmpathicAct$, which may not be reached by looking at isolated dimensions, e.g., Act, alone. Hence, the ontology can provide different explanations of the users' perceptions by considering ramification effects (Section 5.4) captured in the model.

In this section, we have presented an evaluation of the developed model of computational empathy, applied as a tool to measure perceptions of empathy in two interactive agents, Wysa and Replika. We show how we can understand and explain the perceptions of empathy with (1) an analysis considering average measures on the main dimensions, (2) a detailed bottom-up analysis considering possibilistic intervals on each sub-dimension, and (3) a semantic analysis considering ramification effects between interdependent concepts.

7. Discussion

Through an examination of prior research on empathy and conversational agents in the domain of health and well-being, along with a subsequent user study exploring human-agent interaction, we have identified a set of necessary conditions for computational empathy. These conditions can be encapsulated in a formal, multi-dimensional definition and semantic model—an ontology—consisting of six overarching dimensions, expanded in more explicit underlying concepts. The hierarchical structure of the formal definition aids in understanding empathy at different levels of abstraction, allowing to incorporate perspectives from various human empathy theories. This includes the two routes to empathy, affective and cognitive, from a psychological standpoint (Goldman, 2011), the observed relations between perception and action in neuroscience (Preston, 2007), and concepts such as shared emotions (Szanto & Krueger, 2019) and social bonds (Watt, 2005) from a social cognition perspective.

7.1. Conceptual alignment

We have highlighted the hierarchical structure of the ontology to provide a bottom-up approach for defining abstract empathy concepts. Moreover, we see the potential of integrating external ontologies to define sub-concepts of computational empathy. For instance, there are ontologies developed to define emotions (López, Gil, García, Cearreta, & Garay, 2008), personality (Alamsyah, Dudija, & Widiyanesti, 2021), honesty (Wester, Brännström, Nieves, & Van Berkel, 2023), trust (Viljanen, 2005), and other factors that may be important for empathy in particular settings. Nevertheless, when integrating external ontologies, careful attention must be paid to ensure alignment of the concepts. The issue of "alignment" is a well-recognized challenge in ontology integration (Kolli, 2023; de Souza & Davis, 2004; Zimmermann & Le Duc, 2008) and must be addressed to maintain semantic consistency and meaningful integration. The potential misalignment of concepts highlights the importance of maintaining the hierarchical structure within our proposed ontology, as it facilitates the clear definition of concepts in lower levels. The hierarchical structure supports in achieving conceptual alignment from multiple perspectives, whether it is between systems through integrated ontologies or in relation to human comprehension. When eliciting perceptions from human participants, it is crucial for our observations to resonate with the participants' subjective understanding of the concepts. By employing explicit low-level concepts, a human-system conceptual alignment can be achieved, ensuring that our inquiries connect with the participants' understanding while maintaining the intended meaning of ontological concepts.

7.2. Standardization and data interoperability

The primary aim of this work has been to present a high-level general definition of computational empathy that, through adaptable underlying sub-concepts, serves as a versatile tool applicable to various systems and applications. A significant advantage of the presented ontology is that it promotes data interoperability (Paolucci & Souville, 2012) and standardization (Gal & Rubinfeld, 2019). By establishing a shared concept of empathy among various systems, we enable a common ground for communication and understanding. In the rapidly evolving landscape of communicative systems, having standardized concepts becomes relevant to ensure effective interaction and knowledge sharing. Hence, our intention is to establish a conceptual foundation that can be adapted and specialized for particular agents as needed. The user study involving the chatbots, Replika and Wysa, serves as an illustrative example of the potential applications of the definition.

7.3. Specalizations of the model

The identified high level dimensions of computational empathy can be understood in different ways for different user groups, contexts and agents. A general perspective may exclude important dimensions and/or overemphasize others. Thus, the model must be specialized for particular use-cases. For example, empathy dimensions relevant in chatbots may be different from the dimensions relevant in autonomous cars or smart homes. Moreover, the model can be analyzed for understanding which dimensions are relevant for perceived empathy in humans. Through extended user studies analyzed though quantitative methods, we can identify dominant dimensions using a weighting equation: $E = \sum_{i=1}^{n}(d_i \cdot w_i)$, such that E represents the overall empathy score, $\sum$ denotes the summation over all empathy dimensions (from d_1 to d_n), and $w_i \in [0, 1]$ represents a dimension's relevance for a particular population and/or type of interactive agent. Let us observe that w_i is assigned to 1 for all dimensions in the current study. Empirical studies with a substantial sample size are essential for understanding the prominence of specific empathy dimensions within various populations and interactive agents.

We have presented how the approach can be used for assessing empathy. However, the proposed definition can further be applied in automated reasoning, utilized by an agent, for recognizing empathy in differnt kinds of agents it interacts with or observes. For instance: A chatbot may use the tool for assessing its own empathic performance when interacting with humans; A chatbot may in a similar way deliberate about empathy in human users; An observer agent may assess the empathy of a chatbot, a social robot, or two humans interacting. Given the complexity of these diverse scenarios, each case demands a dedicated investigation.

7.4. The empathic machine

Traditionally, a rational software agent is designed following a "sense, think, act" (STA) paradigm (Rao & Georgeff, 1995; Winikoff, Padgham, & Harland, 2001), where an agent perceives the world to update its knowledge, deliberates about its beliefs of the world to decide on its actions, and finally actuates onto the world to fulfill its goals. A notion of "communicate" (C) has been argued to be appended to the paradigm (Siegel, 2003) due to requirements of delivering and updating an agent's knowledge through interactions with other agents. When that communication concerns human agents, a notion of *empathy* comes into play. For building the empathic machine, a formal understanding of empathy is a central for guiding the agent's behavior, for interpreting observations of its interlocutors, and as overarching principles of design. An empathic agent has been suggested to be designed as a rational agent (Alfonso, Vivancos, & Botti, 2017; Brännström & Nieves, 2022; Kampik, Nieves, & Lindgren, 2019; Ochs, Sadek, & Pelachaud, 2012). In accordance, we suggest an empathic rational agent design, where deliberation and means-end reasoning processes are constrained and guided by empathic principles, such that the agent engages in interaction through perception and action by considering theory of mind, internally and externally manifests affective states, and maintains social capabilities through ratification and interpersonal relations.

8. Conclusion and future research

With recent advances in intelligent technology, such as the recent leap in Large Language Models, including GPT (Team OpenAI, 2022), T5 (Ni et al., 2021), and LaMDA (Thoppilan et al., 2022), showcasing remarkable ability to generate human-like language, have led to notable strides in the development of chatbots. Naturally, interactive software agents are increasingly becoming a part of our everyday lives; at home, in mobile devices, and in a range of online services. With new possibilities of processing data and knowledge from social and affective observations of humans, an aim for interactive agents is to be empathic in their interactions with humans. This involves the ability to perceive, interpret and act in a proactive manner that considers human emotions, thoughts, and behaviors. To accomplish such empathic capabilities in machines, a formal understanding of empathy is required. The focus of this study has been to explore and define computational empathy, where the main contributions are: (1) a formal multi-dimensional definition of computational empathy, (2) a qualitative methodology for evaluating computational empathy, specifying an assessment protocol and an analysis method using precise and imprecise uncertainty measures, (3) an OWL ontology implementation, provided as open access material, and (4) a qualitative assessment of computational empathy of two state-of-the-art mental health chatbots, Replika and Wysa.

While computational empathy research is in an early stage (Asada, 2015; Kampik et al., 2019; Lowmanstone, 2021; Paiva et al., 2017; Yalçın, 2019), interactive intelligent agents are progressively deployed in mental health applications that in various degrees demand empathic capabilities to connect with users. With the proposed definition and tool, we can assess the current empathic capacity in the state-of-the-art interactive agents and define aims for the next generation, empathic agents. The introduced ontology for empathy is the first formal and computational ontology that captures functional necessary conditions for computational empathy. The definition is versatile in its application areas, where we see potential for adjusting the model through agent-specific weights, to be utilized for measuring perceptions of empathy in a wide range of interactive agents, such as chatbots, robots, smart homes, autonomous cars, and possibly even humans, despite their varying degrees of transparency. The ontology can further be applied in automated empathy reasoning.

Future research aims to extend the model by exploring weighted empathy dimensions for specific populations or interactive agents, potentially personalizing empathy perceptions. This requires participatory studies with diverse user groups and a span of different interactive agents to acknowledge and manage potential stereotyping of the model. In order to incorporate such dynamics, fuzzy ontology representations (Bobillo & Straccia, 2011; Calegari & Ciucci, 2007; Nagypál & Motik, 2003) can be considered, enabling complex queries by considering imprecise empathy measures. This will enable inferences about imprecise measures to further enhance the model's potential for data interoperability in automated reasoning, sharing empathy assessments and perceptions across human and system boundaries. This can provide enhanced robustness, personalization and trust in human–system interactions. In this direction, we see the potential to cluster the dimensions of computational empathy in terms of requirements for Trustworthy Artificial Intelligence (Dignum, 2019; European Commission and Directorate-General for Communications Networks, Content and Technology, 2019; Vianello, Laine, & Tuomi, 2023; Yeung, 2020), where empathy may have multiple concurrences. Empathic capabilities of intelligent systems, such as in the context of health and well-being, can affect their compliance to different ethical requirements, such as human agency, explainability and robustness. In order to assess the trustworthiness of intelligent systems, an increasing demand is to assess computational empathy.

Declaration of competing interest

The authors declare no conflicts of interest.

Data availability

Data will be made available on request.

Acknowledgments

This work was partially funded by the Knut and Alice Wallenberg Foundation.

References

Ahmad, A., Li, K., Feng, C., Asim, S. M., Yousif, A., & Ge, S. (2018). An empirical study of investigating mobile applications development challenges. *IEEE Access*, *6*, 17711–17728.

Alamsyah, A., Dudija, N., & Widiyanesti, S. (2021). New approach of measuring human personality traits using ontology-based model from social media data. *Information*, *12*(10), 413.

Alfonso, B., Vivancos, E., & Botti, V. (2017). Toward formal modeling of affective agents in a BDI architecture. *ACM Transactions on Internet Technology (TOIT)*, *17*(1), 1–23.

Asada, M. (2015). Towards artificial empathy: how can artificial empathy follow the developmental pathway of natural empathy?. *International Journal of Social Robotics*, *7*, 19–33.

Baader, F., Calvanese, D., McGuinness, D. L., Nardi, D., & Patel-Schneider, P. F. (2003). *The description logic handbook: Theory, implementation, and applications*. Cambridge University Press.

Baader, F., Calvanese, D., McGuinness, D., Patel-Schneider, P., Nardi, D., et al. (2003). *The description logic handbook: Theory, implementation and applications*. Cambridge University Press.

Bae Brandtzæg, P. B., Skjuve, M., Kristoffer Dysthe, K. K., & Følstad, A. (2021). When the social becomes non-human: Young People's perception of social support in chatbots. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–13).

Beilharz, F., Sukunesan, S., Rossell, S. L., Kulkarni, J., Sharp, G., et al. (2021). Development of a positive body image chatbot (KIT) with Young people and parents/carers: Qualitative focus group study. *Journal of Medical Internet Research*, *23*(6), Article e27807.

Beredo, J., Bautista, C. M., Cordel, M., & Ong, E. (2021). Generating empathetic responses with a pre-trained conversational model. In *International Conference on Text, Speech, and Dialogue* (pp. 147–158). Springer.

Bobillo, F., & Straccia, U. (2011). Fuzzy ontology representation using OWL 2. *International Journal of Approximate Reasoning*, *52*(7), 1073–1094.

Boukricha, H., Wachsmuth, I., Carminati, M. N., & Knoeferle, P. (2013). A computational model of empathy: Empirical evaluation. In *2013 Humaine Association conference on affective computing and intelligent interaction* (pp. 1–6). IEEE.

Brandtzaeg, P. B., & Følstad, A. (2018). Chatbots: changing user needs and motivations. *Interactions*, *25*(5), 38–43.

Brännström, A., & Nieves, J. C. (2022). Emotional reasoning in an action language for emotion-aware planning. In *International Conference on Logic Programming and Nonmonotonic Reasoning* (pp. 103–116). Springer.

Burgoon, J. K., Stern, L. A., & Dillman, L. (1995). *Interpersonal adaptation: Dyadic interaction patterns*. Cambridge University Press.

Calegari, S., & Ciucci, D. (2007). Fuzzy ontology, fuzzy description logics and fuzzy-owl. In *International workshop on fuzzy logic and applications* (pp. 118–126). Springer.

Cameron, G., Cameron, D., Megaw, G., Bond, R., Mulvenna, M., O'Neill, S., et al. (2018). Assessing the usability of a chatbot for mental health care. In *International conference on internet science* (pp. 121–132). Springer.

Casas, J., Spring, T., Daher, K., Mugellini, E., Khaled, O. A., & Cudré-Mauroux, P. (2021). Enhancing conversational agents with empathic abilities. In *Proceedings of the 21st ACM international conference on intelligent virtual agents* (pp. 41–47).

Ceha, J., Lee, K. J., Nilsen, E., Goh, J., & Law, E. (2021). Can a humorous conversational agent enhance learning experience and outcomes? In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–14).

Chen, Z., Lu, Y., Nieminen, M. P., & Lucero, A. (2020). Creating a chatbot for and with migrants: Chatbot personality drives co-design activities. In *Proceedings of the 2020 ACM designing interactive systems conference* (pp. 219–230).

Chun Tie, Y., Birks, M., & Francis, K. (2019). Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine*, *7*, Article 2050312118822927.

Chung, K., Cho, H. Y., & Park, J. Y. (2021). A chatbot for perinatal women's and partners' obstetric and mental health care: Development and usability evaluation study. *JMIR Medical Informatics*, *9*(3), Article e18607.

Croes, E. A., & Antheunis, M. L. (2021). Can we be friends with mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships*, *38*(1), 279–300.

De Gennaro, M., Krumhuber, E. G., & Lucas, G. (2020). Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in Psychology*, *10*, 3061.

Decety, J., & Jackson, P. L. (2004). The functional architecture of human empathy. *Behavioral and Cognitive Neuroscience Reviews*, *3*(2), 71–100.

Dignum, V. (2019). *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer Nature.

Duan, J., Zhao, H., Zhou, Q., Qiu, M., & Liu, M. (2020). A study of pre-trained language models in natural language processing. In *2020 IEEE International Conference on Smart Cloud (SmartCloud)* (pp. 116–121). IEEE.

Dubois, D., & Prade, H. (1998). Possibility theory: qualitative and quantitative aspects. In *Quantified representation of uncertainty and imprecision* (pp. 169–226). Springer.

El-Masri, M. M., & Mowbray, F. I. (2019). Data collection, management, entry, and analysis. In *Conducting the DNP project: Practical steps when the proposal is complete*. Springer Publishing Company.

Elliott, R., Bohart, A. C., Watson, J. C., & Greenberg, L. S. (2011). Empathy. *Psychotherapy*, *48*(1), 43.

Elliott, R., Bohart, A. C., Watson, J. C., & Murphy, D. (2018). Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, *55*(4), 399.

European Commission and Directorate-General for Communications Networks, Content and Technology (2019). *Ethics guidelines for trustworthy AI*. Publications Office, http://dx.doi.org/10.2759/177365.

Gabrielli, S., Rizzi, S., Carbone, S., & Donisi, V. (2020). A chatbot-based coaching intervention for adolescents to promote life skills: Pilot study. *JMIR Human Factors*, *7*(1), Article e16762.

Gal, M. S., & Rubinfeld, D. L. (2019). Data standardization. *NYUL Review*, *94*, 737.

Gebhard, P., Aylett, R., Higashinaka, R., Jokinen, K., Tanaka, H., & Yoshino, K. (2021). Modeling trust and empathy for socially interactive robots. In *Multimodal agents for ageing and multicultural societies* (pp. 21–60). Springer.

Ghandeharioun, A., McDuff, D., Czerwinski, M., & Rowan, K. (2019). Towards understanding emotional intelligence for behavior change chatbots. In *2019 8th international conference on affective computing and intelligent interaction* (pp. 8–14). IEEE.

Goldman, A. (2011). Two routes to empathy. *Empathy: Philosophical and Psychological Perspectives*, 31–44.

Greer, S., Ramo, D., Chang, Y. J., Fu, M., Moskowitz, J., & Haritatos, J. (2019). Use of the chatbot "vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. *JMIR mHealth and uHealth*, *7*(10), Article e15018.

Grové, C. (2021). Co-developing a mental health and wellbeing chatbot with and for young people. *Frontiers in Psychiatry*, 1664.

Guthridge, M., & Giummarra, M. J. (2021). The taxonomy of empathy: A meta-definition and the nine dimensions of the empathic system. *Journal of Humanistic Psychology*, Article 00221678211018015.

Hauser-Ulrich, S., Künzli, H., Meier-Peterhans, D., & Kowatsch, T. (2020). A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): pilot randomized controlled trial. *JMIR mHealth and uHealth*, *8*(4), Article e15806.

Hogan, R. (1969). Development of an empathy scale. *Journal of Consulting and Clinical Psychology*, *33*(3), 307.

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, *6*(11), Article e12106.

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018). Evaluating and informing the design of chatbots. In *Proceedings of the 2018 designing interactive systems conference* (pp. 895–906).

Jang, S., Kim, J. J., Kim, S. J., Hong, J., Kim, S., & Kim, E. (2021). Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study. *International Journal of Medical Informatics*, *150*, Article 104440.

Jolliffe, D., & Farrington, D. P. (2006). Development and validation of the basic empathy scale. *Journal of Adolescence*, *29*(4), 589–611.

Kampik, T., Nieves, J. C., & Lindgren, H. (2019). Empathic autonomous agents. In *Engineering Multi-Agent Systems: 6th International Workshop, EMAS 2018, Stockholm, Sweden, July 14-15, 2018, Revised Selected Papers 6* (pp. 181–201). Springer.

Kolli, M. (2023). A bigraphical approach to model and verify ontology alignment. *International Journal of Ad Hoc and Ubiquitous Computing*, *43*(3), 127–143.

Kraus, M., Seldschopf, P., & Minker, W. (2021). Towards the development of a trustworthy chatbot for mental health applications. In *International conference on multimedia modeling* (pp. 354–366). Springer.

Lee, M., Ackermans, S., Van As, N., Chang, H., Lucas, E., & IJsselsteijn, W. (2019). Caring for vincent: a chatbot for self-compassion. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–13).

Lee, Y. C., Yamashita, N., & Huang, Y. (2020). Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1–27.

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., & Dolan, B. (2016). A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Li, Y., Li, K., Ning, H., Xia, X., Guo, Y., Wei, C., et al. (2021). Towards an online empathetic chatbot with emotion causes. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval* (pp. 2041–2045).

Lifschitz, V., Porter, B., & Van Harmelen, F. (2008). *Handbook of knowledge representation*. Elsevier.

Lin, Z., Xu, P., Winata, G. I., Siddique, F. B., Liu, Z., Shin, J., et al. (2019). CAiRE: An empathetic neural chatbot. arXiv preprint arXiv:1907.12108.

López, J. M., Gil, R., García, R., Cearreta, I., & Garay, N. (2008). Towards an ontology for describing emotions. In *Emerging technologies and information systems for the Knowledge Society: First world summit on the Knowledge Society, WSKS 2008, Athens, Greece, September 24-26, 2008. proceedings 1* (pp. 96–104). Springer.

Lowmanstone, L. (2021). *Computational empathy* (Ph.D. thesis), Harvard University.

Luo, T. C., Aguilera, A., Lyles, C. R., & Figueroa, C. A. (2021). Promoting physical activity through conversational agents: mixed methods systematic review. *Journal of Medical Internet Research*, *23*(9), Article e25486.

Ly, K. H., Ly, A. M., & Andersson, G. (2017). A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. *Internet Interventions*, *10*, 39–46.

Maeda, E., Miyata, A., Boivin, J., Nomura, K., Kumazawa, Y., Shirasawa, H., et al. (2020). Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial. *Reproductive BioMedicine Online*, *41*(6), 1133–1143.

Medeiros, L., Gerritsen, C., & Bosse, T. (2019). Towards humanlike chatbots helping users cope with stressful situations. In *International conference on computational collective intelligence* (pp. 232–243). Springer.

Mehrabian, A., & Epstein, N. (1972). A measure of emotional empathy. *Journal of Personality*.

Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. M. (2018). Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *Journal of Medical Internet Research*, *20*(6), Article e10148.

Nagypál, G., & Motik, B. (2003). A fuzzy model for representing uncertain, subjective, and vague temporal knowledge in ontologies. In *OTM confederated international conferences" on the move to meaningful internet systems"* (pp. 906–923). Springer.

Ni, J., Ábrego, G. H., Constant, N., Ma, J., Hall, K. B., Cer, D., et al. (2021). Sentence-t5: scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*.

Ochs, M., Sadek, D., & Pelachaud, C. (2012). A formal model of emotions for an empathic rational dialog agent. *Autonomous Agents and Multi-Agent Systems*, *24*, 410–440.

Paiva, A., Leite, I., Boukricha, H., & Wachsmuth, I. (2017). Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *7*(3), 1–40.

Paolucci, M., & Souville, B. (2012). Data interoperability in the future of middleware. *Journal of Internet Services and Applications*, *3*(1), 127–131.

Pidgeon, N. F., Turner, B. A., & Blockley, D. I. (1991). The use of grounded theory for conceptual analysis in knowledge elicitation. *International Journal of Man-Machine Studies*, *35*(2), 151–173.

Preston, S. D. (2007). A perception-action model for empathy. *Empathy In Mental Illness*, *1*, 428–447.

Rahman, R., Rahman, M. R., Tripto, N. I., Ali, M. E., Apon, S. H., & Shahriyar, R. (2021). AdolescentBot: Understanding opportunities for chatbots in combating adolescent sexual and reproductive health problems in Bangladesh. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–15).

Rao, A. S., & Georgeff, M. (1995). Bdi agents: from theory to practice.. *95*, In *Proceedings of the First International Conference on Multiagent Systems* (pp. 312–319).

Ray, P. P. (2023). Chatgpt: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., et al. (2021). PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. *Systematic Reviews*, *10*(1), 1–19.

Rose-Davis, B., Van Woensel, W., Abidi, S. R., Stringer, E., & Abidi, S. S. R. (2022). Semantic knowledge modeling and evaluation of argument theory to develop dialogue based patient education systems for chronic disease self-management. *International Journal of Medical Informatics*, *160*, Article 104693.

Ryu, H., Kim, S., Kim, D., Han, S., Lee, K., & Kang, Y. (2020). Simple and steady interactions win the healthy mentality: Designing a chatbot service for the elderly. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–25.

Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., et al. (2018). Saturation in qualitative research: exploring its conceptualization and operationalization. *Quality & Quantity*, *52*(4), 1893–1907.

Schulman, D., Bickmore, T., & Sidner, C. (2011). An intelligent conversational agent for promoting long-term health behavior change using motivational interviewing. In *2011 AAAI spring symposium series*.

Shang, L., Lu, Z., & Li, H. (2015). Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

Shumanov, M., & Johnson, L. (2021). Making conversations with chatbots more personalized. *Computers in Human Behavior*, *117*, 106627.

Sia, D. E., Yu, M. J., Daliva, J. L., Montenegro, J., & Ong, E. (2021). Investigating the acceptability and perceived effectiveness of a chatbot in helping students assess their well-being. In *Asian CHI symposium 2021* (pp. 34–40).

Siegel, M. (2003). The sense-think-act paradigm revisited. In *1st international workshop on robotic sensing* (pp. 5–pp). IEEE.

de Souza, K. X. S., & Davis, J. (2004). Aligning ontologies and evaluating concept similarities. In *OTM confederated international conferences" on the move to meaningful internet systems"* (pp. 1012–1029). Springer.

Szanto, T., & Krueger, J. (2019). Introduction: empathy, shared emotions, and social identity. *Topoi*, *38*(1), 153–162.

Team OpenAI (2022). *ChatGPT: Optimizing language models for dialogue*.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., et al. (2022). Lamda: language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Vianello, A., Laine, S., & Tuomi, E. (2023). Improving trustworthiness of AI solutions: A qualitative approach to support ethically-grounded AI design. *International Journal of Human-Computer Interaction*, *39*(7), 1405–1422.

Viljanen, L. (2005). Towards an ontology of trust. In *International conference on trust, privacy and security in digital business* (pp. 175–184). Springer.

Wasil, A. R., Palermo, E. H., Lorenzo-Luaces, L., & DeRubeis, R. J. (2021). Is there an app for that? A review of popular apps for depression, anxiety, and well-being. *Cognitive and Behavioral Practice*.

Watt, D. F. (2005). Social bonds and the nature of empathy. *Journal of Consciousness Studies*, *12*(8–9), 185–209.

Wester, J., Brännström, A., Nieves, J. C., & Van Berkel, N. (2023). "You've got a friend in me": a formal understanding of the critical friend agent. In *Proceedings of the 11th International Conference on Human-Agent Interaction* (pp. 443–445).

Winikoff, M., Padgham, L., & Harland, J. (2001). Simplifying the development of intelligent agents. In *Australian joint conference on artificial intelligence* (pp. 557–568). Springer.

Yalçın, Ö. N. (2019). Evaluating empathy in artificial agents. arXiv preprint arXiv: 1908.05341.

Yeung, K. (2020). Recommendation of the council on artificial intelligence (oecd). *International Legal Materials*, *59*(1), 27–34.

Yuen, H. K., & Richards, T. J. (1993). GTKAT: a grounded theory based knowledge acquisition tool for expert systems. In *Proceedings 1993 the first New Zealand international two-stream conference on artificial neural networks and expert systems* (pp. 152–155). IEEE.

Zadeh, L. A. (1978). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, *1*(1), 3–28.

Zhang, P., Wang, J., & Li, R. (2023). Tourism-type ontology framework for tourism-type classification, naming, and knowledge organization. *Heliyon*, *9*(4).

Zimmermann, A., & Le Duc, C. (2008). Reasoning with a network of aligned ontologies. In *International conference on web reasoning and rule systems* (pp. 43–57). Springer.