Journal of Intelligent & Fuzzy Systems xx (20xx) x–xx
DOI:10.3233/JIFS-219336
IOS Press

Digital Companions for Well-being: Challenges and Opportunities

Juan Carlos Nieves[a,*], Mauricio Osorio[b], David Rojas-Velazquez[c], Yazmín Magallanes[b] and Andreas Brännström[a]

[a] *Department of Computing Science, Umeå University, SE-901 87, Umeå, Sweden*
[b] *University of the Americas Puebla, Mexico*
[c] *Division of Pharmacology, Utrecht Institute for Pharmaceutical Sciences, Faculty of Science, Utrecht University, The Netherlands*

Abstract. Humans have evolved to seek social connections, extending beyond interactions with living beings. The digitization of society has led to interactions with non-living entities, such as digital companions, aimed at supporting mental well-being. This literature review surveys the latest developments in digital companions for mental health, employing a hybrid search strategy that identified 67 relevant articles from 2014 to 2022. We identified that by the nature of the digital companions' purposes, it is important to consider person profiles for: a) to generate both person-oriented and empathetic responses from these virtual companions, b) to keep track of the person's conversations, activities, therapy, and progress, and c) to allow portability and compatibility between digital companions. We established a taxonomy for digital companions in the scope of mental well-being. We also identified open challenges in the scope of digital companions related to ethical, technical, and socio-technical points of view. We provided documentation about what these issues mean, and discuss possible alternatives to approach them.

Keywords: Conversational agents, well-being, mental health, trustworthy artificial intelligence

1. Introduction

Mental well-being is described as a state of health and happiness, where individuals realize their potential, successfully navigate life's challenges, participate in productive work, and make meaningful contributions to their community (Cambridge dictionary[1] and the World Health Organization[2]).

The concept of well-being encompasses various dimensions, from subjective well-being [3] and psychological well-being [39] to social well-being [39]. Given its close link to endurance and happiness, it can be approached from both hedonic and eudaimonic perspectives [1]. Moreover, research suggests that well-being is a skill that can be practiced and strengthened, emphasizing the role of compassion and selflessness [9, 29]. Mental well-being activities, particularly for stress, anxiety, and depression management, have benefited from advances in performance, interaction, and processing capacities of interactive and intelligent software agents [17, 35].

A conversational agent or Digital Companion (DC) is a software entity that uses artificial intelligence (AI) techniques to simulate a conversation with a person either by written messages or by voice [17]. These DC have advantages over human healthcare professionals; for instance, they can help with a complex task with efficiency and accuracy. DC that provide well-

[*] Corresponding author. Juan Carlos Nieves, Department of Computing Science, Umeå University, SE-901 87, Umeå, Sweden. E-mail: juan.carlos.nieves@umu.se.

[1] https://dictionary.cambridge.org/es/diccionario/ingles/companion

[2] https://www.who.int/

being care are not susceptible to fatigue, boredom, or burnout, and they are immune to personal biases that human therapists may have [23]. One problem that DC solve is that assistance is available at any time and in any place, as long as the DC is installed on a mobile device[3]. People[4] can receive information about health conditions, self-care counseling, therapeutic activities, among others [23]. DC technology can interact with people with mental disorders. Nevertheless, giving that each person has a lot of private or sensitive information about them, makes us question the information privacy policies that are considered in DC: is there data protection? Are there privacy policies? In addition to the privacy policies, ethical aspects related to the design and evaluation of the DC, and interaction with people and transparency when providing information or therapies must be considered. We have observed that there are surveys that analyze DC for healthcare purposes oriented to review the characteristics, current applications, and evaluation measures [22]; to provide a systematic review of the most influential user studies focused on development, human perception, and interactions [30]; or to conceptualize the scope and to work out the current state of the art of DC fostering mental health [4]. The authors in previous surveys provide reviews related to type of technology (platform supporting the DC), technical evaluation measures (dialog success rate, word accuracy, percentage of words correctly understood, among others), health domain, study types and methods, user experience, technical performance, technical implementation, objectives (improvement of symptoms, support exercises) age, gender, among others [4, 22, 30]. However, these works are not considered ethical issues, privacy policies implementations, transparency, tests, or inclusive aspects. This paper in the format of survey provides a deep overview of AI and well-being from a social and technical perspective by discussing the development of DC, reviewing their architectures, ethical aspects, and exploring strengths and weaknesses for future applications. The contribution of this work is summarized as follows:

- A definition of mental well-being under the context of digital companions.
- A technical analysis of a selection of DC to identify common architectures.
- Highlight the importance of using person profiles, a missing element that is important when designing a DC.
- A classification of the DC according to the design and user experience.
- An ethical analysis in terms of interaction, transparency, tests, and privacy.
- A socio-technical evaluation of commercial digital companions.

The rest of the paper is organized as follows: Section 2 describes the methodology for the analyzed DC selection, Section 3 provides a background of DC for well-being, Section 4 presents a technical analysis for digital companions, Section 5 discusses ethical aspects of well-being and digital companions, Section 6 presents a discussion regarding the challenges in digital companions for mental well-being, and Section 7 concludes the paper.

1. Methods

To conduct this study, we applied a hybrid search strategy combining searching in SCOPUS[5] with backward and forward snowballing. We aimed to identify articles reporting research in DC designed to support well-being. We focused our study on conversational agents that take any unconstrained input.

The review has three stages: identifying relevant studies, study selection, and reporting the results. During the step of identifying relevant studies, two searches were done: a focus topic and another on a string search. With respect to the focus topic search, we explored results in the scope of well-being from a multidisciplinary view.

The string search consisted of the terms: digital companion, well-being, Artificial Intelligence, software, and design (variations and synonyms of the words were also used). The initial search resulted in 67 articles. The title and abstract of each citation were analyzed according to the following predetermined inclusion and exclusion criteria:

1. Articles published in the English language
2. Articles published between 2014 and 2022.
3. Articles that describe applications for supporting well-being.

The results of the string search are mainly presented in Section 2.1 and Section 2.2.

[3] In the context of accessibility and portability without the need to be connected to the internet

[4] In this work, people or/and person are considered as an app user or a clinical patient

[5] https://www.elsevier.com/solutions/scopus/

Table 1
General architecture of digital companions

Reference	Script-based	AI techniques implemented for its operation	Cognitive theory implemented
[8]	No	If-Then system	No
[37]	No	NLP[7]	MOST[8] Model
[10]	Yes	No	PSHHI[9]
[7]	Yes	Intelligent agent	CBT[10], Behavioral Activation
[33]	Yes	No	MISC[11]
[25]	Yes	ECA[12], intelligent agents, Speech recognition, NLP	CBT, the OCC Model[13]
[34]	Yes	Emotions classification	No
[40]	No	Multi-agent system	CBT, Behavioral Activation
[14]	Yes	NLP, Intelligent agent	CBT
[24]	Yes	Intelligent agent, rule-based system	Positive psychology, CBT
[28]	Yes	State Machine	No
[21]	No	Machine Learning, NLP	No
[27]	No	Machine learning and Logistic Regression	No

2. Results

In this section we will present the findings during the analysis of the selected references, an analysis is presented from the technical design, ethical awareness, and socio-technical points of view. In particular, we will present our findings regarding the current development on DCs for well-being.

2.1. Digital companions technical evaluation

In this section, we provide a technical evaluation of different DCs that have been reported in the literature during the last years.

2.1.1. Architecture analysis of DCs

We selected 20 proposals of DCs based on the information provided about their architectures and functionalities. We consider these proposals to represent a general approach when well-being companions are developed. It can be seen in Table 1, most of the proposals are script-based, which means they have preloaded conversations as decision tree, depending on the person's inputs, the companion provides the corresponding answer [7, 10, 14, 24, 25, 28, 33, 34]. The works that do not have this functionality are those with a knowledge base[6], and the companion can decide what answer it can provide to the person through artificial intelligence techniques such as natural language processing (NLP), speech recognition (SR), among others [8, 21, 37, 40].

Digital companions in the well-being domain are considered agents that help people in mental health problems such as stress management, depression management, and anxiety management. This is the reason why the implementation of intelligent agents (IA) is common; some of them just use IA to select questions and answers that will be displayed to the person as plain text in a console [7, 24, 40].

In some cases, these IAs are combined with other artificial intelligence techniques such as IA-NLP [14], or IA-NLP-Speech-Recognition [25]. Other uses IA as an interaction element between companion and person. These agents are called Embodied Conversational Agents (ECA) [25]. Few are the digital companions who do not use IA implementations, instead of that, they use artificial intelligence (AI) techniques such as: if-then systems [8], classification algorithms [34], natural language processing [37], state machine [28], combination between machine learning and NLP [21], or machine learning models with logistic regression for emotional and sentiment analysis to monitor the user's emotion at every step and provides appropriate responses and feedback [27] or no artificial intelligence technique [10, 33].

It was identified that most DC have a cognitive model implemented on which the DC's responses and activities are based. These models were proposed by specialists in mental well-being and are composed of several tasks that the person must perform to work on his/her management of anxiety, depression, or stress. For example, the Moderated Online Social Therapy

[6] A knowledge base (KB) is a database used to store structured and unstructured information used by a computer system.

[7] Natural Language Processing

[8] The Moderated Online Social Therapy

[9] Physiological synchrony found in human-human interaction

[10] Cognitive-Behavior Therapy

[11] Motivational Interviewing Skills Code

[12] Embodied Conversational Agents

[13] Initials of the authors Ortony, Clore, and Collins: the OCC model

Model (the MOST model) is based on information from young people's feedback, research in mental health, and human-computer interaction. The MOST model unites online social media, interactive therapy modules and, peer and professional moderation, creating a constant flow for the person between the social and therapy elements [2]. This model was used in [37].

The physiological synchrony results from human-to-human interaction where the behavior of one individual becomes the stimulus for the other, producing an iterative co-action effect. For example, in yoga breathing exercises, the coach and the trained person enter into the dynamic of influencing each other until a state of relaxation is achieved [10]. This concept is used and modeled to impact the breathing rhythm in the person. In this case, a virtual human starts at a preset breathing rate. The human breathing rate changes gradually until the desired frequency is achieved, causing a state of relaxation in the person [10].

Motivational Interviewing (MI) is a client-centered therapeutic approach in DC architectures that focuses on evoking personal reasons for change and promoting autonomy [38]. It is based on the person's goals and values and emphasizes collaboration and intrinsic motivation [38]. The Motivational Interviewing Skill Code (MISC) assesses MI quality and has various applications, including feedback during MI learning, training effectiveness evaluation, psychotherapy support, and treatment outcome prediction [26]. Both MI and MISC were utilized in [33].

Five of the twelve works analyzed used Cognitive Behavior Therapy (CBT). CBT is a class of interventions that share the premise that cognitive factors maintain mental and psychological disorders. This treatment approach establishes that maladaptive cognitions contribute to the maintenance of emotional distress and behavioral problems. These maladaptive cognitions include beliefs or schemes about the world, the self, and the future, which give rise to specific and automatic thoughts in particular situations. The basic model states that therapeutic strategies to change these cognitions lead to changes in emotional distress and problem behaviors [20]. In CBT therapy, the person becomes an active participant in a collaborative problem-solving process to test and challenge the validity of maladaptive cognitions and modify behavior patterns. CBT refers to a set of interventions that combine cognitive, behavioral, and emotion-focused techniques [20]. An example of the implementation of this model in a DC is in [14].

Some authors combine CBT with other therapeutic models such as behavioral activation (BA) as can be seen in [7, 40]. BA is a structured psychosocial approach based on behavior change that aims to alleviate mental disorders such as depression, anxiety, or stress and avoid relapses. BA has the premise that the problems in the lives of vulnerable people and their behavioral responses to such problems, reduce their ability to experience positive rewards from their environment. The treatment aims to systematically increase activation in ways that help patients experience greater contact with reward sources in their lives and solve life problems. The treatment focus on activation and on processes that inhibit activation, such as escape and avoidance behaviors, to increase experiences that are pleasurable or productive and improve life context [12].

Another combination with CBT is with the psychological model of emotions proposed by Orthony, Clore, and Collins in 1990 (the OCC Model) to provide an emotional reaction to ECA [25]. The OCC Model is a well known psychological model for emotions that describes the cognitive processes to generate human emotions based on environment evaluations carried out by people [31]. In [24], authors combined CBT with positive psychology, which is the study of the processes that contribute to the optimal functioning of people, groups, and institutions [16].

In the absence of cognitive models, some DCs function as information search engines [8]. Others employ classification techniques to analyze user inputs, assess mental and emotional states, and suggest treatment recommendations [24]. In [28], a chat-bot was created to support in-person relationships using machine states. Alternatively, some DCs utilize support vector machine, naive Bayes, and natural language processing to correlate input messages with stored text, providing output messages [21]. In [27], machine learning techniques are used to analyze text-emotion relationships for tracking user emotions and tailoring responses.

2.1.2. Psychological profiles

One area of opportunity that we identified during the technical analysis is the use of person profiles. These profiles are an essential tool for tracking a person's progress during physical, medicated, or mental treatment. In particular, we highlight two types of person profiles: 1) psychological profile and 2) emotional profile. We do not consider aspects such as age, gender, and/or ethnicity because the analyzed works consider it during test of both usability and results.

The psychological profile is a biographical sketch based on the behavior of a person [41] so that psychologists have detailed information on each patient and can design personalized help plans using CBT, Behavioral Activation (BA), Positive Psychology, the MOST model, among others. **The emotional profile** is the collection of information about a person's emotional characteristics, personality, behavior, and interests to better understand why a person behaves a certain way. In the well-being domain, it has been seen that negative emotions (when they are extreme, prolonged, or contextually inappropriate) can generate problems for individuals and society such as anxiety, depression, or stress. The intervention strategies that cultivate positive emotions are particularly suited for preventing and treating these problems [15]. Techniques like relaxation training, finding positive meaning, invoking empathy, amusement, or interest can generate positive emotions and relieve negative emotions [15]. Therefore, it is crucial to design personalized activities to create positive emotions, and this is possible by having the person's emotional information documented in the same way as in psychological therapy. Hence the importance of managing person profiles in help tools such as DC.

When conducting technical analysis, the lack of use of person profiles was noted. This being an important element for monitoring, proposing well-being tasks, and for interaction with people. We consider that both emotional and psychological profiles are important missing elements in the DC architecture for the domain of mental well-being. As can be seen in Table 2, only four DC [7, 10, 24, 25] implements psychological profiles in their architecture, these profiles are (most of the time) a history of previous interactions with the person. What is most striking is that **the DCs do not take users' emotion into consideration when providing interventions** to help stress, anxiety, or depression management. This is the reason why proposing the implementation and use of both emotional and psychological profiles in the DC architectures could improve the efficiency in techniques, interaction, engagement, and results obtained from DC for well-being.

It is essential to mention that there are commercial products that serve as digital companions for mental well-being: Replika[14], Yana[15], and Wysa[16]. These digital companions lack detailed architecture and AI technique descriptions, excluding them from the technical analysis. Nevertheless, they are considered in this work as special case studies only from both an ethical and socio-technical perspective.

Table 2
Digital companions that consider person profiles

Reference	Psychological profile
[8]	No
[37]	No
[10]	Yes
[7]	Yes
[33]	No
[25]	Yes
[34]	No
[40]	No
[14]	No
[24]	Yes
[28]	No
[21]	No
[27]	No

2.2. *Ethical awareness and a Socio-technical evaluation*

In this section, we provide the Ethical Awareness and Socio-technical Assessment of DCs reported in selected references as well as commercial DCs.

2.2.1. *Ethical awareness*

Ethical awareness has emerged as a fundamental requirement in Information Systems that have a degree of autonomy in their decision processes and aim to interact with humans [11, 19]. One can highlight that digital companions are a particular class of systems that need to be aware of the social implications of their actions and interactions with a persona who requires support to cope with his or her mental state.

In the current state of art, there is no classification on the expected ethical awareness on digital companions. Nevertheless, one can identify a classification on AI-based systems that was introduced by Dignum ([11]). This classification is depicted by Fig. 2. The classification suggested by Dignum aims to cluster intelligent autonomous systems and the expected social awareness level about them. By taking the perception of the end-users on digital companions, one can also classify them as *tools*, *assistants*, and *partners*. By tools, we means systems that are only reactive to the requests of a persona. Assistants are systems that are reactive and also have a level proactiveness in their services. Partner digital companions

[14] https://replika.ai/

[15] https://yana.com.mx/

[16] https://www.wysa.io/

can be envisioned as systems that can act close to human capabilities of interaction, see Fig. 1. In these three classes of digital companions, one can expected an increasing level of autonomy as it is suggested by Fig. 2. By observing Table 3, one can notice that current digital companions are mainly perceived and designed as either tools or assistants. From this view, one can expect at least a functional level of ethics awareness in the decisions that are taken by current digital companions. However, we have observed that ethical principles are only considered in the settings of data governance. This means that current digital companions are mainly concerned on be law compliance with regulations such as the General Data Protection Regulation [13].

The High-Level Expert Group on AI of EU presented the Ethics Guidelines for Trustworthy Artificial Intelligence [19]. These guidelines had fostered seven key requirements to envision trustworthy AI systems: [P1] Human agency and oversight; [P2] Technical Robustness and safety; [P3] Privacy and data governance; [P4] Transparency; [P5] Diversity, non-discrimination and fairness; [P6] Societal and environmental well-being; and [P7] Accountability.

The seven principles, including P1, P4, and P6, serve as essential criteria for trustworthy digital companions. However, current digital companions lack substantial evidence of trustworthiness, particularly in areas like transparency and explainability [18]. Many digital companions function as opaque "black boxes", with limited skills to explain their decisions. Achieving greater explainability relies on the development and integration of explainable algorithms into digital companion architectures.

2.2.2. Socio-technical evaluation

We have observed that there is a lack of methods and tools to evaluate digital companions from a socio-technical point of view. In this section, we suggest an approach to assess digital companions from a social-technical point of view. To exemplify the suggested evaluation approach, we have selected and evaluated four applications focused on well-being. Our evaluation has 13 criteria based and adjusted from the framework proposed by [42], who exemplify their usefulness in some contact-tracking applications to fight COVID-19. We used these criteria and assessed well-being companions' apps on a scale from 0 to 2 in a qualitative way, where generally 2 means that the application complies with the requirement, 1 means that the application partially complies with the condition, and 0 means that there is no evidence regarding the need. We use the following criteria to assess the selected applications.

1. Respecting fundamental rights of individuals: This comprises the rights to safety, health, and nondiscrimination (2). Partially including these rights (1). There is no information (0).
2. Privacy and data protection: Clearly defines the app's purpose and assess its usage mechanism (2). Partially including this information (1). There is no information (0).
3. Transparency rights: Ensure users are notified, have control over their data, and disclose the personal data collected (2). Partially containing this information (1). There is no information (0).
4. Avoid discrimination: The app needs to prevent stigmatization (2). Partially including this information (1). There is no information (0).
5. Accessibility: Possibility to be used by all regardless of demographics, language, disability, digital literacy, and financial accessibility (2). Partially including this information (1). There is no information (0).
6. Education and tutorials: Ensure that people are informed and capable of using the app correctly, including in-app help (2). External materials, such as a website (1). There is no help (0).
7. Data management: Ensure that only data strictly necessary are processed (2). Partially including this information (1). There is no information (0).
8. Security: Person authentication to prevent risks such as access, modification, or disclosure of the data (2). Partially including this information (1). There is no information (0).
9. Application easy to deactivate/remove: It has clear instructions or automatically initiates the removal process on request (2). Difficulties for removing the app and the data (1). The information can't be removed (0).
10. Open-source code: Participatory and multidisciplinary development (2). Access to open-source code without the possibility of contributing (1). Non-open source (0).
11. Public ownership: Ownership by State (2). Health agency or research institute (1). Private or commercial party (0).
12. Legislation and policy: Include a legal framework (2). Partial government policy (1). There is no information (0).
13. Design Impact Assessment and Open Development Process: Explicit design process, including aims and motivation, stakeholders,

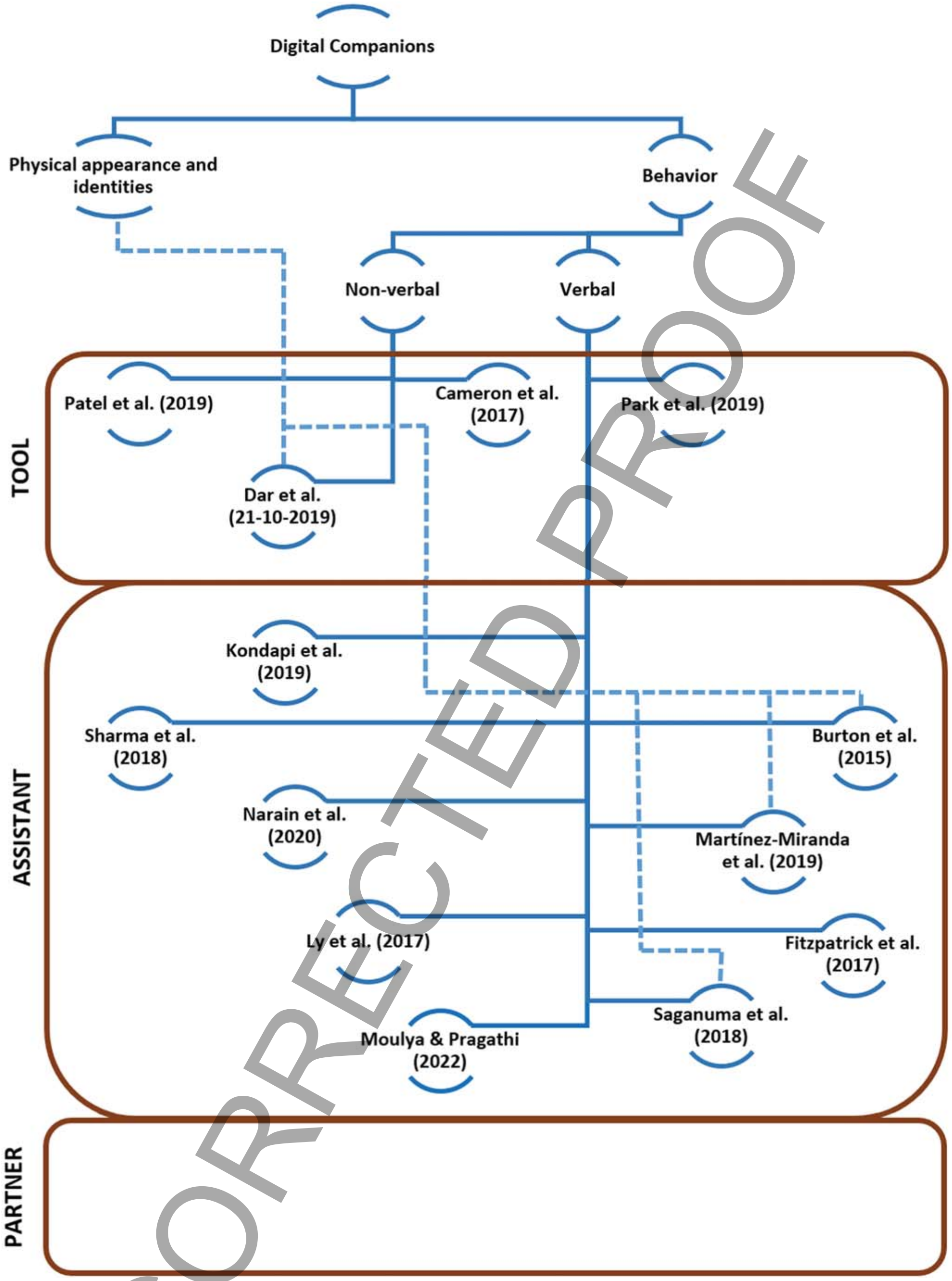

Fig. 1. Taxonomy proposed for digital companions for well-being.

and impact assessment (2). Partially including this information (1). There is no information (0).

Figure 3 shows the results after evaluating Woebot, Replika, Yana, and Wisa (our selected applications). We observe that all the applications have low scores regarding accessibility, public ownership, design impact assessment, and are not developed under open-source code. Most of the applications lack tutorials about their usage and policies regarding individuals' fundamental rights and discrimination.

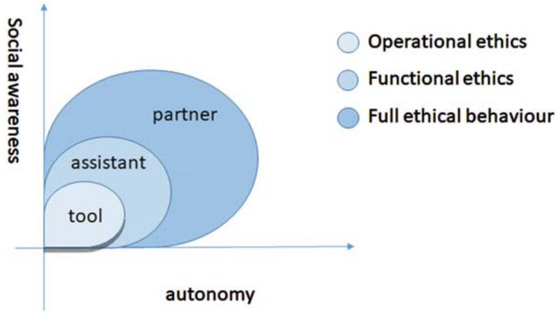

Fig. 2. Autonomy and Social-Awareness in AI-based systems [11].

3. Discussion

Well-being can be considered a skill that can be improved and tools like DC can be helpful in that improvement. Despite this, there are still sensitive issues for development, design, and testing that must be considered. In this work, we identify several challenges in the DC for mental well-being such as:

- Ethical Concerns: Commercial products consider ethical aspects related to testing with humans, information privacy, and data management, but they lack transparency, and there's no option for users to inquire about the DC's responses.
- Lack of Psychological Profiles: The absence of psychological and emotional profiles hinders personalized responses and can lead to tedious interactions.
- Limited Clinical Evaluation: DCs lack comprehensive clinical evaluations, making it challenging to assess their impact on individuals' well-being and the therapies they provide.
- Documentation Challenges: While psychological models are documented, their practical implementation and comparability with other DCs remain unclear, hindering research and improvement.

The above challenges appear to be easy to solve, but a multidisciplinary effort is needed to generate standards for the evaluation, testing protocols, development processes, documentation of ethical and transparency policies for the DC for mental well-being. In this work, we tried to provide some guidelines for the evaluation of DC from the ethical and socio-technical point of view. These guidelines could be extended, shortened, or modified to be improved and applied in DC in the well-being

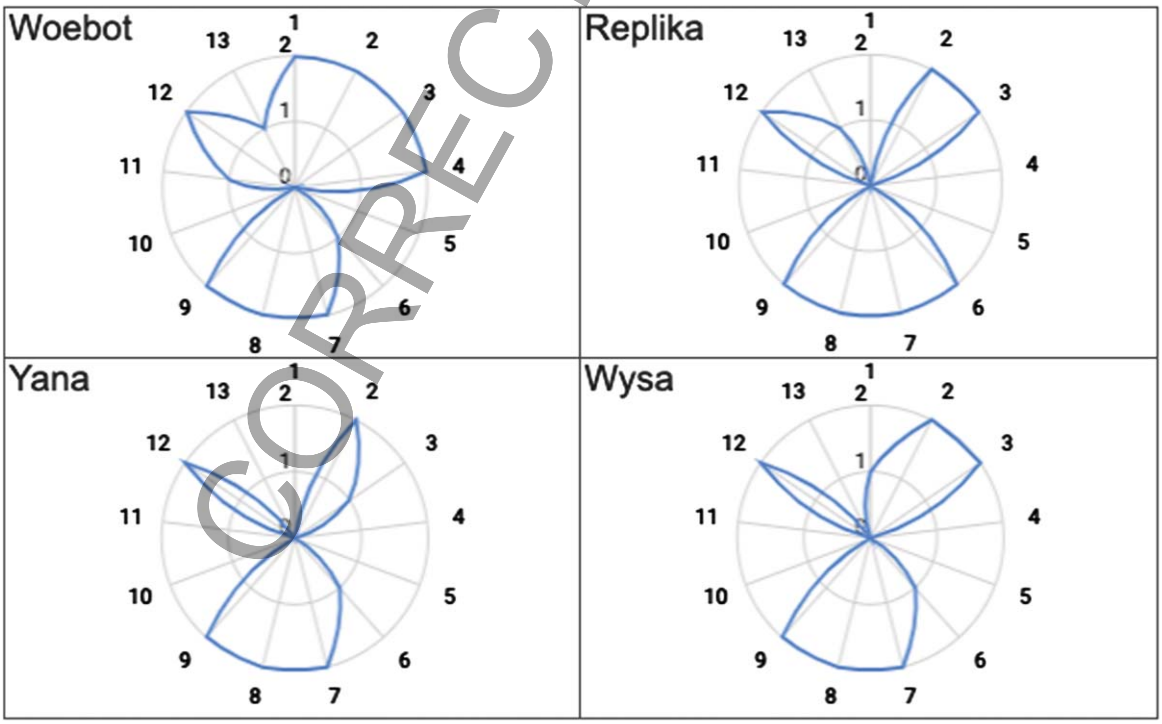

Fig. 3. Applications evaluated through a socio-technical framework. The numbers represent each of the criteria and its compliance.

Table 3
Ethical awareness of digital companions

Reference	Perception of the end-users about DC	Perception of the DC based on its design	Ethical principles
[8]	Not specified	Tool	In terms of privacy and security
[37]	Not specified	Assistant	No
[10]	Not specified	Tool	No
[7]	Assistant	Assistant	Complete informed consent
[33]	Tool	Tool	[17]
[25]	Assistant	Assistant	[18]
[34]	Not specified	Tool	No
[40]	Not specified	Assistant	[19]
[14]	Assistant	Assistant	[20]
[24]	Assistant	Assistant	[21]
[28]	Assistant	Assistant	No
[21]	Not specified	Assistant	No
[27]	Not specified	Assistant	No

domain. The technical analysis allows us to identify the absence of cognitive theories oriented to generate a general model of the person's mind and the option for the person to ask about why the DC suggests certain activity or provided some particular response. This theory of mind could represent a general approach to model a person's profile, and both emotional and psychological profiles could be the individual approach. With the implementation of these elements, the DC will have a complete profile of each person and can provide personalized help, activities, and responses. Looking further, if the structure of these profiles is standardized, portability could be accomplished allowing the people to use more than one DC in the process to improve mental well-being. A related kind of standardization have recently been approached in the setting of empathy in interactive agents [6], where "computational empathy" is formally defined and implemented in Web Ontology Language (OWL). Another example, is an implementation of a theory of mind model provided in [36], where the authors present a Multi-Agents architecture based on the theory of mind to model deception reasoning using a mathematical logic language. And one example of the use of profiles and theory of mind for specific purposes can be seen in [32] where the authors complement the theory of mind in [36] with an emotion model and profiles to foreseeing deception. Moreover, in the realm of emotions, a formal model has been introduced to represent the dynamic nature of emotions by formalizing cognitive theories in an action language [5]. This model captures the actions of agents and their consequential impacts on emotions. In the context of DC, these concepts could be used to generate person-oriented help, activities, and responses provided by DC. The integration of all concepts mentioned could lead to DC for mental well-being improvements.

4. Conclusions

In this work, an ethical, socio-technical, and technical analysis was presented. The findings in each analysis carried out allowed us to identify the weaknesses and strengths of DC for mental well-being. We identified that there is complexity and many considerations to define the concept of well-being, that is why in this work we followed a definition for mental well-being (given in the introduction) including Richard Davidson considerations. We identified the common architecture of the DC documented in the literature selected following a hybrid search. This common architecture consists in: a) a psychological basis for providing therapies, task, information, or interventions, b) artificial intelligence techniques (agent-based systems, speech recognition, classification, among others) for the interaction with the person, and c) a method for the generation and continuity of the dialogue between the DC and the person. We also identified the open challenges related to the

[17] It was approved by the Seoul National University Institutional Review Board (IRB No. 1708/001-018).

[18] Ethical standards of the institutional/national research committee and with the 1964 Helsinki declaration.

[19] The authors sought and gained ethics approval from our University Ethics Review Committee.

[20] Stanford School of Medicine's Institutional Review Board.

[21] The institutional review board approved the study protocol. Written informed consent was obtained.

ethical, technical, and socio-technical point of view, we provided documentation about what these issues mean, and discuss possible alternatives to solve them. We highlighted the importance of considering the ethical aspects in the DC for mental well-being since they are not considered in the development of the DC and not at the time of making analyzes and evaluations in other surveys. We described the missing elements in the DC architectures and provided a classification of DC according to people's perception. The intention of this work is to provide the basis to develop a robust, ethical, and trustworthy DC for mental well-being.

This work may serve as a guide for developing a reliable and ethical DC for mental well-being. By addressing challenges like ethical concerns, lack of psychological profiles, limited clinical evaluation, and documentation issues, the paper offers practical insights. The proposed multidisciplinary approach provides a roadmap for creating a more personalized and responsible DC system. In essence, this work provides valuable considerations for building a trustworthy DC that prioritizes ethical standards and user well-being.

References

[1] Aaron Ahuvia, Neil Thin, Dan Haybron, Robert Biswas-Diener, Mathieu Ricard and Jean Timsit, Happiness: An interactionist perspective, *International Journal of Wellbeing* **5**(1) (2015).

[2] Mario Alvarez-Jimenez, J.F. Gleeson, S. Rice, C. Gonzalez-Blanch and S. Bendall, Online peer-to-peer support in youth mental health: seizing the opportunity, *Epidemiology and Psychiatric Sciences* **25**(2) (2016), 123.

[3] Paul Anand, Happiness, well-being and human development: The case for subjective measures. Technical report, United Nations Development Programme, New York, USA, 2016.

[4] Eileen Bendig, Benjamin Erb, Lea Schulze-Thuesing and Harald Baumeister, The next generation: Chatbots in clinical psychology and psychotherapy to foster mental health – a scoping review. Verhaltenstherapie, pages 1–13, 2019.

[5] Andreas Brannstrom and Juan Carlos Nieves, Emotional reasoning in an action language for emotion-aware planning. In International Conference on Logic Programming and Nonmonotonic Reasoning, pages 103–116. Springer, 2022.

[6] Andreas Brannstrom, Joel Wester and Juan Carlos Nieves, A formal understanding of computational empathy in interactive agents, *Cognitive Systems Research* (2023), 101203.

[7] Christopher Burton, Aurora Szentagotai Tatar, Brian McKinstry, Colin Matheson, Silviu Matu, Ramona Moldovan, Michele Macnab, Elaine Farrow, Daniel David, Claudia Pagliari, Antoni Serrano Blanco and MariaWolters, Pilot randomised controlled trial of help4mood, an embodied virtual agent-based system to support treatment of depression, *Journal of Telemedicine and Telecare* **22**(4) (2015), 348–355.

[8] Gillian Cameron, David Cameron, Gavin Megaw, Raymond Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour and Michael McTear, Towards a chatbot for digital counselling. In *Proceedings of the 31st International BCS Human Computer Interaction Conference (HCI 2017)*, pages 1–7. ACM, 2017.

[9] Michaël Dambrun, Matthieu Ricard, Gérard Després, Emilie Drelon, Eva Gibelin, Marion Gibelin, Mélanie Loubeyre, Delphine Py, Aurore Delpy, Céline Garibbo, et al. Measuring happiness: From fluctuating happiness to authentic–durable happiness, *Frontiers in Psychology* **3** (2012), 16. ISSN 1664-1078. doi: 10.3389/fpsyg.2012.00016. URL https://www.frontiersin.org/article/10.3389/fpsyg.2012.00016.

[10] Sanobar Dar, Victoria Lush and Ulysses Bernardet, The virtual human breathing relaxation system. In 2019 5th Experiment International Conference (exp. At'19), pages 276–277. IEEE, 21-10-2019.

[11] Virginia Dignum, Responsable Artificial Intelligence. Springer, 2019. ISBN 978-3-030-30370-9.

[12] Sona Dimidjian, Christopher R. Martell, Michael E. Addis and Ruth Herman-Dunn, Behavioral activation for depression. In David H. Barlow, editor, *Clinical Handbook of Psychological Disorders*, pages 328–364. The Guilford Press, 2014.

[13] European-Parliament, Regulation (eu) 2016/679 – general data protection regulation. *Official Journal of the European Union*, 2018.

[14] Kathleen Kara Fitzpatrick, Alison Darcy and Molly Vierhile, Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial, *JMIR Mental Health* **4**(2) (2017), e19.

[15] Barbara L. Fredrickson, Cultivating positive emotions to optimize health and well-being, *Prevention & Treatment* **3**(1) (2000), 1a.

[16] Shelly L. Gable and Jonathan Haidt, What (and why) is positive psychology? *Review of General Psychology* **9**(2) (2005), 103–110.

[17] Hannah Gaffney, Mansell Warren and Sara Tai, Conversational agents in the treatment of mental health problems: Mixed-method systematic review, *JMIR Mental Health* **6**(10): ConverAgents, 2019.

[18] David Gunning, Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* **2**(2) (2017).

[19] HLEG. Ethics guidelines for trustworthy ai. Technical report, European Commission, 2019.

[20] Stefan G. Hofmann, Anu Asnaani, Imke J.J. Vonk, Alice T. Sawyer and Angela Fang, The efficacy of cognitive behavioral therapy: A review of meta-analyses, *Cognitive Therapy and Research* **36**(5) (2012), 427–440.

[21] Ramya Kondapi, Rahul Kumar Katta and Sirisha Potluri, Pacifiurr: An android chatbot application for human interaction, *International Journal of Recent Technology and Engineering* **7**(5S2) (2019), 101–104.

[22] Liliana Laranjo, Adam G. Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Suriana, Blanca Gallego, Farah Magrabi, Annie Y.S. Lau and Enrico Coiera, Conversational agents in healthcare: a systematic review, *Journal of the American Medical Informatics Association* **25**(9) (2018), 1248–1258.

[23] David D. Luxton, *An Introduction to Artificial Intelligence in Behavioral and Mental Health Care*, chapter 1, pages 1–26. Academic Press, 2016.

[24] Kien Hoa Ly, Ann-Marie Ly and Gerhard Andersson, A fully automated conversational agent for promoting mental wellbeing: A pilot rct using mixed methods, *Internet Interventions* **10** (2017), 39–46.

[25] Juan Martınez-Miranda, Ariadna Martınez, Roberto Ramos, Héctor Aguilar, Liliana Jiménez, Hodwar Arias, Giovanni Rosales and Elizabeth Valencia, Assessment of users' acceptability of a mobile-based embodied conversational agent for the prevention and detection of suicidal behaviour, *Journal of Medical Systems* **43**(8) (2019), 246.

[26] William R. Miller, Theresa B. Moyers, Denise Ernst and Paul Amrhein, Manual for the motivational interviewing skill code (misc), *Center on Alcoholism, Substance Abuse, and Addictions*, The University of New Mexico, 2003.

[27] S. Moulya and T.R. Pragathi, Mental health assist and diagnosis conversational interface using logistic regression model for emotion and sentiment analysis. In *Journal of Physics: Conference Series*, volume 2161, page 012039. IOP Publishing, 2022.

[28] Jaya Narain, Tina Quach, Monique Davey, Hae Won Park, Cynthia Breazeal and Rosalind Picard, Promoting wellbeing with sunny, a chatbot that facilitates positive messages within social groups. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–8. ACM, 2020.

[29] Kristin D. Neff and Emma Seppälä, Compassion, well-being, and the hypo-egoic self. Oxford handbook of hypo-egoic phenomena: Theory and research on the quiet ego, pages 189–202, 2016.

[30] Nahal Norouzi, Kangsoo Kim, Jason Hochreiter, Myungho Lee, Salam Daher, Gerd Bruder and GregWelch, A systematic survey of 15 years of user studies published in the intelligent virtual agents conference. In *Proceedings of the 18th international conference on intelligent virtual agents*, pages 17–22. ACM, 2018.

[31] Andrew Ortony, Gerald L. Clore and Allan Collins, *The cognitive structure of emotions*. Cambridge University Press, 1990.

[32] Mauricio Osorio, Luis Angel Montiel, David Rojas-Velazquez and Juan Carlos Nieves, E-friend: A logical-based ai agent system chat-bot for emotional well-being and mental health, *Deceptive AI*, pages 87–104, 2020.

[33] SoHyun Park, Jeewon Choi, Sungwoo Lee, Changhoon Oh, Changdai Kim, Soohyun La, Joonhwan Lee and Bongwon Suh, Designing a chatbot for a brief motivational interview on stress management: Qualitative case study, *Journal of Medical Internet Research* **21**(4) (2019), e12231.

[34] Falguni Patel, Riya Thakore, Ishita Nandwani and Santosh Kumar Bharti, Combating depression in students using an intelligent chatbot: A cognitive behavioral therapy. In *2019 IEEE 16th India Council International Conference (INDICON)*, pages 1–4. IEEE, 2019.

[35] Davidson Richard, The four keys to well being, *Greater Good Magazine Berkeley*, (03), 2016.

[36] Stefan Sarkadi, Alison R. Panisson, Rafael H. Bordini, Peter McBurney, Simon Parsons and Martin Chapman, Modelling deception using theory of mind in multi-agent systems, *AI Communications* **32**(4) (2019), 287–302.

[37] Bhuvan Sharma, Harshita Puri and Deepika Rawat, Digital psychiatry – curbing depression using therapy chatbot and depression analysis. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pages 627–631. IEEE, 2018.

[38] Rebecca M. Shingleton and Tibor P. Palfai, Technologydelivered adaptations of motivational interviewing for healthrelated behaviors: A systematic review of the current research, *Patient Education and Counseling* **99**(1) (2016), 17–35.

[39] Dana Sidorová, *Well-being, flow experience and personal characteristics of individuals who do extreme sports as serious leisure*. PhD thesis, Masaryk University. Brno, Czech Republic. Czechia, 2015.

[40] Shinichiro Suganuma, Daisuke Sakamoto and Haruhiko Shimoyama, An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: Feasibility and acceptability pilot trial, *JMIR Mental Health* **5**(3) (2018), e10454.

[41] Jennifer Trager and JoAnne Brewster, The effectiveness of psychological profiles, *Journal of Police and Criminal Psychology* **16**(1) (2001), 20–28.

[42] Ricardo Vinuesa, Andreas Theodorou, Manuela Battaglini and Virginia Dignum, A socio-technical framework for digital contact tracing, *Results in Engineering* **8** (2020), 100163. ISSN 2590-1230. doi: https://doi.org/10.1016/j.rineng.2020.100163. URL http://www.sciencedirect.com/science/article/pii/S2590123020300694.