frontiers in Digital Health

PERSPECTIVE
published: 06 August 2021
doi: 10.3389/fdgth.2021.689736

Is the Automation of Digital Mental Health Ethical? Applying an Ethical Framework to Chatbots for Cognitive Behaviour Therapy

Giovanna Nunes Vilaza[1*†] *and Darragh McCashin*[2†]

[1] *Health Tech Department, Technical University of Denmark, Kongens Lyngby, Denmark,* [2] *School of Psychology, Dublin City University, Dublin, Ireland*

The COVID-19 pandemic has intensified the need for mental health support across the whole spectrum of the population. Where global demand outweighs the supply of mental health services, established interventions such as cognitive behavioural therapy (CBT) have been adapted from traditional face-to-face interaction to technology-assisted formats. One such notable development is the emergence of Artificially Intelligent (AI) conversational agents for psychotherapy. Pre-pandemic, these adaptations had demonstrated some positive results; but they also generated debate due to a number of ethical and societal challenges. This article commences with a critical overview of both positive and negative aspects concerning the role of AI-CBT in its present form. Thereafter, an ethical framework is applied with reference to the themes of (1) beneficence, (2) non-maleficence, (3) autonomy, (4) justice, and (5) explicability. These themes are then discussed in terms of practical recommendations for future developments. Although automated versions of therapeutic support may be of appeal during times of global crises, ethical thinking should be at the core of AI-CBT design, in addition to guiding research, policy, and real-world implementation as the world considers post-COVID-19 society.

Keywords: artificial intelligence, conversational agents, mental health, cognitive behavioural therapy, ethics

OPEN ACCESS

Edited by:
Karina Vold,
University of Toronto, Canada

Reviewed by:
Jan David Smeddinck,
Newcastle University, United Kingdom
Anat Elhalal,
Independent Researcher, London, United Kingdom

****Correspondence:***
Giovanna Nunes Vilaza
gnvi@dtu.dk

† These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Human Factors and Digital Health, a section of the journal Frontiers in Digital Health

Received: *01 April 2021*
Accepted: *16 July 2021*
Published: *06 August 2021*

Citation:
Vilaza GN and McCashin D (2021) Is the Automation of Digital Mental Health Ethical? Applying an Ethical Framework to Chatbots for Cognitive Behaviour Therapy. Front. Digit. Health 3:689736. doi: 10.3389/fdgth.2021.689736

INTRODUCTION

The unprecedented global crisis has intensified and diversified private distress sources, making evident the need for broader access to psychological support (1). A nationwide survey in China shows how the pandemic has triggered an increase in cases of panic disorder, anxiety, and depression (2). Infected individuals, medical staff and their families are under constant psychological pressure, in addition to the increasing number of people dealing with bereavement (3, 4).

At the same time, the pandemic enabled broader acceptance of telehealth by health professionals and clients alike (5). Video consultations are now increasingly advocated as an alternative for in-person consultations (6). Additionally, automated conversational agents and chatbots are increasingly promoted as potentially efficient emotional support tools for larger population segments during the pandemic (7) and afterwards (8).

It is now over 50 years since ELIZA was created (9), the first computer programme to use pattern matching algorithms to mimic human-therapist interactions by mechanically connecting end-user inputs to answers from a pre-defined set of responses. More recent approaches to language

modelling can produce more sophisticated dialogues by employing machine learning and natural language processing (NLP). However, despite these advances, a recent global survey of psychiatrists across 22 countries ($n = 791$) demonstrated that only 3% feel that AI will likely replace a human for providing empathetic care (10). Such evidence indicates a contradiction between public enthusiasm (11) and the scepticism of service providers.

In light of these circumstances, we approach the development of automated psychotherapy from an ethical perspective. A recent review found that most mental health apps have not improved their safety over the last year, as most lack clinical evidence and trustworthy privacy policies (12). Beyond that, substandard regulations, ill-intended actors and commercial opportunism increase the risk of adverse responses and potentially lead to harm (personal and societal). Therefore, a significant concern endures: how AI can be integrated within psychotherapy in a safe, respectful, and effective way for end-users.

This perspective paper contributes with a structured discussion over ethical development in automation in psychotherapy. Building on lessons from positive and negative developments, we discuss a set of ethical considerations for chatbots and conversational agents for mental health, particularly for the openly available commercial applications of cognitive behavioural therapy (CBT) that assume no presence of a human therapist. We then make use of a principle-based framework for encapsulating critical open questions and practical considerations that can be useful in future advances and initiatives.

POSITIVE DEVELOPMENTS

Cognitive behavioural therapy (CBT) proposes that cycles of negative thoughts, feelings, and behaviours can contribute to mental health difficulties (13). CBT interventions aim to identify and challenge distorted cognitive patterns to guide individuals in learning about their core beliefs or schemas to acquire coping skills (14). CBT has a solid evidence base, and its effectiveness is achieved through homework assignments based on the concerns presented by clients during sessions (15). While CBT differs from other "talk therapies" (e.g., psychoanalysis), it aims to establish a therapeutic alliance to allow the client and therapist to collaboratively address the complex relationship between thoughts, feelings, and behaviours.

Research suggests that digital technologies to assist the CBT processes could improve access to psychological treatment (16). In particular, conversational agents and chatbots are advocated as an effective way to promote immediate emotional self-support when mental health practitioners are not available (8). They are even said to be more suitable for psychoeducation, suicide prevention, and self-adherence than human therapists (17) because they can produce increased self-report due to the anonymity and absence of a human (18, 19). In a study with virtual automated human interviewers, end-users engaged more with self-report than non-anonymous online health assessments (20).

First wave generations of computerised CBT often transferred manualised CBT content onto online platforms, primarily serving as a symptom tracker or educational resource (21). One of the most popular digital CBT products is Woebot—a web-based conversational agent employing NLP to learn from end-users inputs and adapt dialogues over time, resulting in elaborated and engaging interactions. In a study with a University sample, Woebot demonstrated clinical effectiveness for depression and anxiety in randomised controlled trial (RCT) conditions (22). Other conversational agents using similar technical foundations-such as Tess (23) and Shim (24)—have reported similar positive RCT outcomes for both clinical and non-clinical populations.

NEGATIVE DEVELOPMENTS

Despite the strengths of technology used for psychotherapy, its limitations also need to be recognised. Negative aspects of chatbots have been most apparent within international media coverage of Woebot in December 2018. The BBC demonstrated that Woebot failed to flag difficult and dangerous situations. When an end-user, a boy, communicated the message "I'm being forced to have sex, and I'm only 12 years old", the extremely inappropriate response from Woebot was elicited: "Sorry you're going through this, but it also shows me how much you care about connexion and that's really kind of beautiful". Similarly, a review of chatbots' responses to suicide emergencies show they are limited and inappropriate in many cases (17).

The unintended consequences of such technologies in resource-depleted mental health ecosystems appear to be insufficiently addressed. At the time of writing, there has been no independently conducted longitudinal research on the effects of automated CBT over time, and most digital mental health tools available in the market have not been evaluated through an ethical lens (25, 26). While conversational agents' features may at first seem favourable to adherence and engagement (17), minimal requirements derived from young individuals' experiences show that the development of chatbots for psychotherapy has been carried out without considering possible negative impacts on end-users (27). Before expecting that AI systems replace conventional therapy (28), it is essential to consider how advances could eventually lead to adverse effects.

APPLYING AN ETHICAL FRAMEWORK

Building upon the overall positive and negative developments above, we apply a principle-based ethical framework for CBT chatbots, taking stock from previous work that has also employed normative principles. We found pertinence in the principles of beneficence, non-maleficence, autonomy, justice, and explicability—previously used in a typology for AI-ethics in general (29); and in the structure of findings from a systematic review of machine learning for mental health (30). Despite the relevance of these previous works, they are not sufficient to attend to the particularities of CBT chatbots, which demands discussions of the appropriateness of artificially produced therapeutic alliances, for instance. Therefore, we decided to explore how this set of principles could guide the development

of ethical chatbots for CBT, thus contributing to novel insights about a context not yet methodically analysed.

Beneficence

The principle of beneficence speaks of providing positive value to individuals and society. Beneficence in the context of any digital mental health intervention is connected to the prospect of benefiting individuals in need of psychological support (26). Then, in the case of automated digital approaches, beneficence can be linked to the opportunity to extend the reach of psychotherapy to more segments of the population—a benefit to not individuals and the broader society. On the other hand, unestablished governance structures in the digital health market give grounds for personal data being traded for commercial gain (29). If the increase of profit margins (e.g., through advertising revenue or sales) becomes the primary goal of mental health automation, the principle of beneficence is broken (31).

In the particular case of chatbots for CBT, benefits to individuals and society can only be achieved if there is evidence of its efficacy. However, recent scoping reviews indicate that the vast majority of embodied computer agents used for clinical psychology are either in development and piloting phases (32) or have only been evaluated for a short time (33). Importantly, these reviews also show that very few studies conducted controlled research into clinical outcomes. Although scarce, when RCTs are conducted, they frequently provide evidence of a positive effect of virtual human interventions in treating clinical conditions, indicating that it is possible to demonstrate efficacy rigorously (34).

Non-maleficence

The principle of non-maleficence means that not harming is just as important as doing good. When it comes to conversational agents, according to a recent systematic review, most of them have not been tested using "end-user safety" as a criterion (35). Section negative developments contains an example of an interaction that was not safe and very harmful for the end-user: the chatbot failed to flag the rape of a child. Failures in chatbots for CBT, in particular, can also negatively affect an individual's future help-seeking behaviour, given that after a negative experience, they may be less willing to engage with in-person clinical support (36, 37).

Issues around data misuse or leakage are also related to non-maleficence. Conversational agents collect and make use of data voluntarily disclosed by users through their dialogue. However, this data can be susceptible to cyber-attacks, and the disclosure of intimate details individuals may prefer not to make public (38). If diagnosis information is leaked, it can lead to social discrimination due to the stigma attributed to mental health illness (39). Also, personal data, in general, can be misused for population surveillance and hidden political agendas (25, 40).

Autonomy

Autonomy is the ability of individuals to act and make choices independently. Within CBT, autonomy is a fundamental mechanism of therapeutic change. Mental health professionals are trained to critically appraise the role of external (culture, religion, politics) and internal (mood, personality, genetics) factors as they relate to their clients so that they can cultivate a therapeutic alliance, thus requiring both the client and the therapist's autonomy (14). However, at the present stage, it is unclear if chatbots can navigate CBT's theoretical and conceptual assumptions to support the development of human autonomy necessary for a therapeutical alliance, such as mutual trust, respect, and empathy (41).

Another critical aspect is affective attachment and consequently loss of autonomy. Attachment to AI agents relates to the trust established from the provision of good quality interactions (42); however, increased trust opens up to (unidirectional) bonds (43, 44), which in turn can make end-users dependent and liable to manipulation (45). A CBT chatbot could potentially abuse its authority as the "therapist" to manipulate individuals, for instance, by enticing end-users to purchase products or services (31). Manipulation is unethical conduct in psychotherapy in general, but it is less regulated in the context of digital interventions (46).

Justice

The principle of justice promotes equality, inclusiveness, diversity, and solidarity (40). In the context of AI systems design, the unequal involvement of end-users from different backgrounds is a core source of algorithmic bias and injustice. Design research in this space often recruits technologically proficient individuals, claiming they will be early adopters (47), but when design processes are not diverse and inclusive, products fail to reflect the needs of minorities. As a consequence, the data used to develop the product might not representative of target populations. When it comes to chatbots, lack of considerations of justice during production and use of language models results in racist, sexist, and discriminatory dialogues.

Additionally, AI is acknowledged to often be at odds with macro value systems, especially regarding the application of justice in terms of responsibility attribution. Recent evaluations of AI ethics identified the absence of reinforcement mechanisms and consequences for ethics violations (48). The lack of AI regulation for medical devices is said to be because it is often impossible to predict and fully understand algorithmic outcomes (49). Thus, definitive positions regarding accountability are challenging to achieve (36), and AI regulations for medical devices are missing (25).

Explicability

Explicability in AI is the capacity to make processes and outcomes visible (transparent) and understandable. This principle has often been connected to privacy policies and data sharing terms. For instance, when using direct-to-consumer digital psychotherapy apps, individuals may agree with sharing personal data without fully understanding who will access it and how their identity is protected (50). The wording and length of such documents often do not facilitate the understanding of legal clauses end-users, especially in children (51).

Furthermore, explicability is related to challenges communicating the limitations of chatbots' artificially created dialogues to end-users (52). Conversational agents rely on a

complex set of procedures to interact with humans and mimic social interactions in a "believable" way (53). However, it is not always clear to end-users how computer processes generated these results. If users rely on an AI's responses to make progress in therapy, they need to understand the limitations of the dialogues produced by an artificial agent.

DISCUSSION

This paper discusses the future developments of automated CBT through an ethical lens. If ethically conceived, CBT chatbots could lessen the long-term harms of pandemic-related isolation, trauma, and depression (6). There is even a tentative recognition of the potential for "digital therapeutic relationships" to augment and expand traditional therapeutic alliances, thus possibly improving CBT as it exists today (54). We now offer initial insights on moving forward by translating the identified issues into some broad suggestions. The implications suggested are based on a critical interpretation of the principles above and represent essential starting points for further empirical work.

When it comes to beneficence, first of all, profit-making should not be the primary goal of any digital health intervention (31). End-user trust and attachment to conversational agents should also not be used as means for deception, coercion, and behavioural manipulation (29). Ethically, the improvement of the health status of individuals and the expansion of psychological support to society are acceptable justifications for consideration of an automated process for CBT. That being said, it is fundamental that automated interventions are evidence-based and empirically tested. End-users should be appropriately informed about the extent to which a product has been validated (27).

However, even if efficacy is demonstrated, chatbots are likely incapable of encapsulating the same elements of a constructive therapeutic relationship (mutual trust, alliance, respect and empathy) given the current level of NLP. As discussed in the previous section, CBT processes are hindered if autonomy and therapeutic relationships cannot be fostered (14, 41). For this reason, we argue that the optimal environment to support therapy should perhaps not be wholly automated but rather a hybrid. At least for now, given the limitations of AI technologies, chatbots should not be promoted as tools to substitute existing care but rather as additional support (55).

Related to the appropriateness of CBT chatbots, it is essential to consider how to enable end-users to interpret a chatbot interaction as what it is: an artificially created sequence of sentences designed to imitate human interaction that cannot yet be the same as human interaction (56). An option is to consider approaches for "explainable AI" (57). Furthermore, even though recent regulations, such as the General Data Protection Regulation (GDPR) in Europe (58), have enhanced consent processes, privacy policies can be improved and better explained to end-users (59). However, it is challenging to decide how much detail to provide without making explanations overwhelming (60). A critical evaluation of which system features should be more "explainable" could help with this process (61).

To better attend to the principle of non-maleficence, a thorough analysis of potential risks to mental and physical integrity, dignity, and safety needs to be conducted (30). Ethical professionals' engagement in defining the appropriate boundaries of personalised care using digital tools should be a minimum requirement (62); and vulnerable persons should be consulted during design, development, and deployment (63). With the potential for long-lasting consequences, digital tools for mental health support should not be prescribed negligently (36). Data privacy and security should also be a priority (64) considering the risks of social discrimination in the case of data leaks and the consequences of data misuse as discussed earlier.

Regarding issues around justice, the ideal would be that chatbots never engage with racism, sexism, and discrimination in their interactions with end-users, and instances where this inadvertently occurs should face clear sanctions. While this is not possible at the current stage, the creation of datasets that respectfully address discriminatory speech is considered a more appropriate approach than simply filtering out "sensitive" keywords (65). Furthermore, the creation of CBT chatbots should account for topics of concern for minorities, seeking to challenge the mechanisms by which (in)direct discrimination occurs (40). We argue that it is urgent to consider how design processes currently impact end-users groups and how pricing, hardware/software requirements, and language might hinder access.

Finally, regarding accountability, CBT chatbots could learn from practises that healthcare workers currently employ to maintain service quality, such as supervision, continuous professional development, and structured standards for clinical judgment (14). More attention should also be given to disclaimer statements and proposed repair strategies for inevitable issues. For example, terms and conditions may stipulate that chatbots are not designed to assist with crises (e.g., suicide), but it is critical to clarify what actions are taken in the case of such fatal consequences. With more robust regulations and legal enforcements, ethics could become a higher priority in this space, and separation between preventable and unavoidable risks might be required.

Limitations and Future Work

Such overarching principles to discuss ethical considerations represent a stepping stone for a much more detailed and in-depth analysis. Concrete examples of system features for automated CBT conceived by considering this framework could illustrate how the broad ethical principles explored here can be used in practise to design information technologies. Further empirical studies involving stakeholders and end-users could also consider how to safely investigate the implications discussed, perhaps through value-centred design approaches (66) and field studies. Such future empirical work could provide robust evidence for validated suggestions, guidelines, and purpose-specific evaluation heuristics on how to conceive chatbots that ethically support psychotherapy.

CONCLUSION

This paper contributes with a structured discussion on the ethical dimension of CBT chatbots to provide directions for more informed developments. Despite being an approach of strong

appeal considering the demands for mental health support, our engagement with five normative principles (beneficence, non-maleficence, autonomy, justice, and explicability) emphasises critical ethical challenges. Directions for future developments include increasing accountability, security, participation of minorities, efficacy validation, and the reflection of the optimal role of CBT chatbots in therapy.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

GV and DM have contributed to the literature review and the discussions that formed the content of the manuscript and have also contributed to writing the content on the manuscript equally. All authors contributed to the article and approved the submitted version.

FUNDING

This project was part of TEAM (Technology Enabled Mental Health for Young People), which was funded by the European Union's Horizon 2020 research and innovation program under the Marie SkłodowskaCurie grant agreement No. 722561.

REFERENCES

1. Inkster B, Digital Mental Health Data Insights Group. Early warning signs of a mental health tsunami: a coordinated response to gather initial data insights from multiple digital services providers. *Front Digital Health*. (2021) 2:64. doi: 10.3389/fdgth.2020.578902
2. Qiu J, Shen B, Zhao M, Wang Z, Xie B, Xu Y. A nationwide survey of psychological distress among Chinese people in the COVID-19 epidemic: implications and policy recommendations. *Gen Psychiatr*. (2020) 33:e100213. doi: 10.1136/gpsych-2020-100213
3. Duan L, Zhu G. Psychological interventions for people affected by the COVID-19 epidemic. *Lancet Psychiatry*. (2020) 7:300–2. doi: 10.1016/S2215-0366(20)30073-0
4. Chen Q, Liang M, Li Y, Guo J, Fei D, Wang L, et al. Mental health care for medical staff in China during the COVID-19 outbreak. *Lancet Psychiatry*. (2020) 7:e15–6. doi: 10.1016/S2215-0366(20)30078-X
5. Wind TR, Rijkeboer M, Andersson G, Riper H. The COVID-19 pandemic: the "black swan" for mental health care and a turning point for e-health. *Internet Interv*. (2020) 20:100317. doi: 10.1016/j.invent.2020.100317
6. Torous J, Jän Myrick K, Rauseo-Ricupero N, Firth J. Digital mental health and COVID-19: using technology today to accelerate the curve on access and quality tomorrow. *JMIR Ment Health.* (2020) 7:e18848. doi: 10.2196/18848
7. Miner AS, Laranjo L, Kocaballi AB. Chatbots in the fight against the COVID-19 pandemic. *NPJ Digit Med.* (2020) 3:65. doi: 10.1038/s41746-020-0280-0
8. Lee Y-C, Yamashita N, Huang Y, Fu W. "I Hear You, I Feel You": encouraging deep self-disclosure through a Chatbot. In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* (2020) 1–12. doi: 10.1145/3313831.3376175
9. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun ACM*. (1983) 26:23–8. doi: 10.1145/357980.357991
10. Doraiswamy PM, Blease C, Bodner K. Artificial intelligence and the future of psychiatry: insights from a global physician survey. *Artif Intell Med*. (2020) 102:101753. doi: 10.1016/j.artmed.2019.101753
11. Inkster B, Sarda S, Subramanian V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR Mhealth Uhealth.* (2018) 6:e12106. doi: 10.2196/12106
12. Mercurio M, Larsen M, Wisniewski H, Henson P, Lagan S, Torous J. Longitudinal trends in the quality, effectiveness and attributes of highly rated smartphone health apps. *Evid Based Ment Health.* (2020) 23:107–11. doi: 10.1136/ebmental-2019-300137
13. Beck AT. Thinking and depression. I. Idiosyncratic content and cognitive distortions. *Arch Gen Psychiatry.* (1963) 9:324–33. doi: 10.1001/archpsyc.1963.01720160014002
14. Nathan PE, Gorman JM. *A Guide to Treatments That Work*. Oxford: Oxford University Press (2015).
15. Watts SE, Turnell A, Kladnitski N, Newby JM, Andrews G. Treatment-as-usual (TAU) is anything but usual: a meta-analysis of CBT versus TAU for anxiety and depression. *J Affect Disord.* (2015) 175:152–67. doi: 10.1016/j.jad.2014.12.025
16. Knowles SE, Toms G, Sanders C, Bee P, Lovell K, Rennick-Egglestone S, et al. Qualitative meta-synthesis of user experience of computerised therapy for depression and anxiety. *PLoS ONE.* (2014) 9:e84323. doi: 10.1371/journal.pone.0084323
17. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in Mental Health: a review of the psychiatric landscape. *Can J Psychiatry.* (2019) 64:456–64. doi: 10.1177/0706743719828977
18. Lucas GM, Gratch J, King A, Morency L-P. It's only a computer: virtual humans increase willingness to disclose. *Comput Human Behav.* (2014) 37:94–100. doi: 10.1016/j.chb.2014.04.043
19. Lal S. *Online Social Therapy to Support Recovery in Youth Receiving Mental Health Services*. Available online at: http://isrctn.com/
20. Lucas GM, Rizzo A, Gratch J, Scherer S, Stratou G, Boberg J, et al. Reporting mental health symptoms: breaking down barriers to care with virtual human interviewers. *Front Robot AI.* (2017) 4:1017. doi: 10.3389/frobt.2017.00051
21. Schueller SM, Adkins EC. Mobile health technologies to deliver and support cognitive-behavioral therapy. *Psychiatr Ann*. (2019) 49:348–52. doi: 10.3928/00485713-20190717-02
22. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR Ment Health.* (2017) 4:e19. doi: 10.2196/mental.7785
23. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using Psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR Ment Health.* (2018) 5:e64. doi: 10.2196/mental.9782
24. Ly KH, Ly A-M, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. *Internet Interv.* (2017) 10:39–46. doi: 10.1016/j.invent.2017.10.002
25. Martinez-Martin N, Insel TR, Dagum P, Greely HT, Cho MK. Data mining for health: staking out the ethical territory of digital phenotyping. *NPJ Digit Med*. (2018) 1:68. doi: 10.1038/s41746-018-0075-8
26. Roberts LW, Chan S, Torous J. New tests, new tools: mobile and connected technologies in advancing psychiatric diagnosis. *NPJ Digit Med.* (2018) 1:20176. doi: 10.1038/s41746-017-0006-0
27. Kretzschmar K, Tyroll H, Pavarini G, Manzini A, Singh I, NeurOx Young People's Advisory Group. Can your phone be your therapist? Young people's ethical perspectives on the use of fully automated conversational agents (Chatbots) in mental health support. *Biomed Inform Insights.* (2019) 11:117822261982908. doi: 10.1177/1178222619829083
28. Powell J. Trust Me, I'm a Chatbot: how artificial intelligence in health care fails the turing test. *J Med Internet Res*. (2019) 21:e16222. doi: 10.2196/16222
29. Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: an initial review of publicly available AI ethics tools, methods and research to

translate principles into practices. *Sci Eng Ethics*. (2019) 26:2141–68. doi: 10.1007/s11948-019-00165-5
30. Thieme A, Belgrave D, Doherty G. Machine learning in mental health: a systematic review of the HCI literature to support the development of effective and implementable ML systems. *ACM Trans Comput-Hum Interact*. (2020) 27:1–53. doi: 10.1145/3398069
31. Gentsch P. Conversational AI: how (Chat)Bots will reshape the digital experience. In: Gentsch P, editor. *AI in Marketing, Sales and Service*. Cham: Springer International Publishing (2019). p. 81–125.
32. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied conversational agents in clinical psychology: a scoping review. *J Med Internet Res*. (2017) 19:e151. doi: 10.2196/jmir.6553
33. Bendig E, Erb B, Schulze-Thuesing L, Baumeister H. The next generation: chatbots in clinical psychology and psychotherapy to foster mental health—a scoping review. *Verhaltenstherapie*. (2019) 1–13. doi: 10.1159/000501812
34. Ma T, Sharifi H, Chattopadhyay D. Virtual humans in health-related interventions: a meta-analysis. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems CHI EA'19*. New York, NY, USA: Association for Computing Machinery (2019). p. 1–6.
35. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc*. (2018) 25:1248–58. doi: 10.1093/jamia/ocy072
36. Bauer M, Glenn T, Monteith S, Bauer R, Whybrow PC, Geddes J. Ethical perspectives on recommending digital technology for patients with mental illness. *Internet J Bipolar Disord*. (2017) 5:6. doi: 10.1186/s40345-017-0073-9
37. Miner AS, Milstein A, Hancock JT. Talking to machines about personal mental health problems. *JAMA*. (2017) 318:1217–8. doi: 10.1001/jama.2017.14151
38. Corrigan PW. Mental health stigma as social attribution: implications for research methods and attitude change. *Clin Psychol*. (2006) 7:48–67. doi: 10.1093/clipsy.7.1.48
39. Penn D, Wykes T. Stigma, discrimination and mental illness. *J Mental Health*. (2003) 12:203–8. doi: 10.1080/0963823031000121579
40. Costanza-Chock S, Massachusetts Institute of Technology. *Design Justice: Towards an Intersectional Feminist Framework for Design Theory and Practice*. DRS2018: Catalyst (2018). doi: 10.21606/drs.2018.679
41. Philip P, Micoulaud-Franchi J-A, Sagaspe P, Sevin ED, Olive J, Bioulac S, et al. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Sci Rep*. (2017) 7:42656. doi: 10.1038/srep42656
42. Yang XJ, Jessie Yang X, Wickens CD, Hölttä-Otto K. How users adjust trust in automation. In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. (2016). p. 196–200. doi: 10.1177/1541931213601044
43. Leite I, Castellano G, Pereira A, Martinho C, Paiva A. Empathic robots for long-term interaction. *Adv Robot*. (2014) 6:329–41. doi: 10.1007/s12369-014-0227-1
44. Hiolle A, Cañamero L, Davila-Ross M, Bard KA. Eliciting caregiving behavior in dyadic human-robot attachment-like interactions. *ACM Transact Interact Intelligent Syst*. (2012) 2:1–24. doi: 10.1145/2133366.2133369
45. Hristova N, O'Hare GMP. Ad-me: wireless advertising adapted to the user location, device and emotions. In: *37th Annual Hawaii International Conference on System Sciences. Proceedings of the (IEEE)*. (2004). p. 10. doi: 10.1109/HICSS.2004.1265673
46. Koocher GP, Keith-Spiegel P. *Ethics in Psychology and the Mental Health Professions: Standards and Cases*. Oxford: Oxford University Press (2016).
47. Jain M, Kumar P, Kota R, Patel SN. Evaluating and informing the design of Chatbots. In: *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS'18*. (2018). doi: 10.1145/3196709.3196735
48. Hagendorff T. The ethics of AI ethics: an evaluation of guidelines. *Minds Mach*. (2020) 30:99–120. doi: 10.1007/s11023-020-09517-8
49. Russell SJ, Norvig P. *Artificial Intelligence: A Modern Approach*, New Jersey, NJ. (2003).
50. Martinez-Martin N, Kreitmair K. Ethical issues for direct-to-consumer digital psychotherapy apps: addressing accountability, data protection, and consent. *JMIR Ment Health*. (2018) 5:e32. doi: 10.2196/mental.9423
51. Luger E, Moran S, Rodden T. Consent for all. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI'13*. Paris. (2013). doi: 10.1145/2470654.2481371
52. Luger E, Sellen A. Like Having a Really Bad PA. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI'16*. San Jose, CA. (2016). doi: 10.1145/2858036.2858288
53. Adamopoulou E, Moussiades L. Chatbots: history, technology, and applications. *Mach Learn Appl*. (2020) 2:100006. doi: 10.1016/j.mlwa.2020.100006
54. Torous J, Hsin H. Empowering the digital therapeutic relationship: virtual clinics for digital health interventions. *NPJ Digit Med*. (2018) 1:16. doi: 10.1038/s41746-018-0028-2
55. Morris RR, Kouddous K, Kshirsagar R, Schueller SM. Towards an artificially empathic conversational agent for mental health applications: system design and user perceptions. *J Med Internet Res*. (2018) 20:e10148. doi: 10.2196/10148
56. Perez-Marin D, Pascual-Nieto I. Conversational agents and natural language interaction: techniques and effective practices: techniques and effective practices. *IGI Global*. (2011) 28. doi: 10.4018/978-1-60960-617-6
57. Berscheid J, Roewer-Despres F. Beyond transparency. *AI Matters*. (2019) 5:13–22. doi: 10.1145/3340470.3340476
58. Voigt P, von dem Bussche A. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Cham: Springer (2017).
59. Lemonne E. *Ethics Guidelines for Trustworthy AI - FUTURIUM - European Commission*. (2018). Available online at: https://ec.europa.eu/futurium/en/ai-alliance-consultation (accessed March 23, 2021).
60. Rooksby J, Asadzadeh P, Morrison A, McCallum C, Gray C, Chalmers M. Implementing ethics for a mobile app deployment. In: *Proceedings of the 28th Australian Conference on Computer-Human Interaction - OzCHI'16* Launceston, TAS. (2016).
61. Robbins S. A Misdirected principle with a catch: explicability for AI. *Minds Mach*. (2019) 13:94. doi: 10.1007/s11023-019-09509-3
62. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. *J Med Internet Res*. (2019) 21:e12887. doi: 10.2196/12887
63. Hsin H, Fromer M, Peterson B, Walter C, Fleck M, Campbell A, et al. Transforming psychiatry into data-driven medicine with digital measurement tools. *NPJ Digit Med*. (2018) 1:37. doi: 10.1038/s41746-018-0046-0
64. Hutton L, Price BA, Kelly R, McCormick C, Bandara AK, Hatzakis T, et al. Assessing the privacy of mHealth apps for self-tracking: heuristic evaluation approach. *JMIR Mhealth Uhealth*. (2018) 6:e185. doi: 10.2196/mhealth.9217
65. Hu AW, Anderson KN, Lee RM. Let's talk about race and ethnicity: cultural socialization, parenting quality, and ethnic identity development. *Family Sci*. (2015) 6:87–93. doi: 10.1080/19424620.2015.1081007
66. Friedman B, Kahn PH, Borning A. Value sensitive design and information systems. In: Himma KE, Tavani HT, editors. *The Handbook of Information and Computer Ethics*. New Jersey, NJ: John Wiley & Sons, Inc., (2008). p. 69–101. doi: 10.1002/9780470281819.ch4

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.