Position paper presented at CONVERSATIONS 2019 - an international workshop on chatbot research, November19-20, Amsterdam, the Netherlands.

Digital Confessions: Exploring the Role of Chatbots in Self-Disclosure

Chris van der Lee, Emmelyn Croes, Jan de Wit, and
Marjolijn Antheunis

Tilburg University, Tilburg School of Humanities and Digital Sciences, Department of Communication and Cognition
`{c.vdrlee, e.a.j.croes, j.m.s.dewit, m.l.antheunis}@tilburguniversity.edu`

Abstract. e-Mental health applications may provide a solution for understaffing issues on the workers' side as well as issues regarding help-seeking (e.g. stigma, high costs) on the patients' side in the mental healthcare domain. Especially the use of conversational AI is seen as a promising solution for these issues. While initial research in this area shows promising results, there is still a lack of empirical research, which makes it difficult to estimate the contribution that conversational agents may have in mental healthcare. The success of these applications may depend on the capabilities of the techniques as well as the degree in which the applications fulfill our goals and expectations. The current paper discusses important challenges that conversational agents may be facing in the mental healthcare domain. Furthermore, PRIESTESS – a newly developed chatbot – is introduced. This chatbot can be used for self-disclosure research to provide information about current beliefs, values, and needs regarding conversational agents which may aid the development of better chatbots.

Keywords: Chatbot · Self-Disclosure · Mental Healthcare · Communication and Technology · Conversational Agent · Conversational AI.

1 Introduction

Mental health clinics are increasingly understaffed. The Netherlands is facing a structural deficit of psychiatrists, psychologists and nurses, which may lead to clinics (temporarily) closing their doors for new patients in the future [14]. Simultaneously, of the roughly 1.9 million adults in the Netherlands that suffer from a psychological disorder (e.g. mood disorder, anxiety disorder, substance disorder), about two thirds does not receive any form of help [30]. These facts illustrate the need for innovations in mental healthcare that decrease the burden on psychiatrists, psychologists and nurses on the one hand, and make it easier for people with a psychological disorder to seek help on the other hand. Investing in e-mental health applications – "mental health services and information delivered or enhanced through the Internet and related technologies" [10, p.3] –

and robotization may help to tackle these issues. Having a conversational agent effectively administer an intervention to a subset of patients through chat, for instance, decreases the workload of mental healthcare professionals, while the relative anonymity, low cost, and easy access of a conversational agent might be appealing to people with a psychological disorder. However, while there are plenty of areas where new mental healthcare technologies may be implemented, the success of these techniques depends on various factors. The current paper discusses important factors that may hamper the usage of e-mental healthcare applications by users, and proposes a focus on self-disclosure research to address these factors. Furthermore, it introduces PRIESTESS: a newly developed open-source conversational agent for eliciting self-disclosure.

2 Conversational agents in the mental healthcare domain

Interest in conversational agents in mental healthcare has increased tremendously in the past few years, as is illustrated by the growing number of new research in this area [5,12,13, for instance]. While the goal of these conversational agents can vary widely, they are most commonly developed for psychoeducational purposes and adherence to treatment [29]. In addition to that, researchers see various other opportunities for conversational agents to support mental healthcare professionals. For instance, chatbots can be deployed for suicide prevention, or to prevent people with psychological problems from relapsing [7,21].

Evaluation studies on using chatbots in the mental healthcare domain show positive results. Conversational agent-based cognitive and behavioral interventions have shown to effectively reduce depression [12,13]. And these outcomes are found to be comparable to therapist-delivered cognitive behavioral therapy [2,4]. However, despite these promising findings, they should be treated with caution. Empirical research on the impact of chatbots in mental health is still rare: a 2014 systematic review of 5,464 papers on internet delivered psychological treatments found that only 5 e-mental health apps were formally evaluated [3]. The existing research is qualitatively heterogeneous in both methodologies and results [29], and most mental health chatbots available for commercial use have not been empirically validated [6].

Research steadily shows that building a relationship with the client is essential for successful therapy [16]. Hence, a chatbot should be able to have appropriate interactions with the client, which implies that it should generate affective communication [25], and that it should be sophisticated and conversation-focused [11]. However, the functionality of most conversational agent applications in this domain is limited: they are more similar to a “choose your own adventure self-help book” [12, p. 3] than to an actual conversational partner.

The success of a therapeutic chatbot lies not only in their ability to respond to a wide variety of requests of the client, but also in being able to establish an emotional connection with the client. The latter is still a challenge for the current generation of chatbots. Though such conversational agents may be feasible soon as automatic information extraction and emotion detection are improving rapidly [19].

3 The role of conversational agents and self-disclosure

One of the key factors that plays a role in mental health interactions, and may help in establishing an emotional connection, is self-disclosure. Research shows that self-disclosure can make people feel relieved [9], even more so when they experience positive reinforcement from their interaction partner [28]. In contrast, people may experience an inhibition to self-disclosure when they feel there are risks involved. For instance, people may be worried that the information they share will be shared with a third party [8] or they may experience a fear of embarrassment or moral judgment [22]. The more people trust their interaction partner, the more they are willing to disclose to them [1].

This paper positions the role of chatbots and self-disclosure in the mental healthcare domain. There are several reasons why chatbots might stimulate self-disclosure, such as accessibility, anonymity, convenience, and their non-judgemental nature [26]. While rare, the studies that have investigated disclosure found very different results: one study found that users disclose as much to chatbots as they would to humans [15], and one found users disclose even more to chatbots [20], while another one found that users disclose less to chatbots and disclose about as much to chatbots as they would in a conventional survey [24]. However, do note that – similar to the research discussed in Section 2 – the way disclosure was measured differed greatly in these studies, and the chatbots were also implemented in vastly different ways.

Of course, what people want to disclose to a chatbot extends beyond the chatbot's mere identity. It also has to do with the content that the conversational agent is producing. For instance, the level of disclosure may have to do with the nature of the questions. Previous findings suggest that relatively neutral questions might not make a difference between chatbots and humans, but users respond with more disclosure intimacy to a chatbot than to a human when they are asked a question that may be embarrassing and might result in negative evaluation [15]. 'Icebreaker questions' (e.g. 'how are you doing?', 'how is the weather?') or humanlike fillers (e.g. 'um', 'ahh') could make the conversational agent appear more humanlike, leading to more affective communication and a sense of shared experience [5], which in turn might increase trust and disclosure intimacy. But such unexpected questions and words may not be compatible with the expected role of the chatbot; or the humanlikeness causes users to expect that the chatbot is also capable of judging them negatively.

Both outcomes might lead to less disclosure intimacy [24,29]. Related to that, users are aware that chatbots cannot emphasize with them the way humans can, so chatbots sharing "lived experiences" (e.g. 'Meditation helped me with my depression') might have an adverse effect on the user-chatbot relationship and therefore affect disclosure intimacy as well [29], while similar content shared by a human conversational partner could be positively received. Empirically investigating the effects of such differences in content produced by the chatbot seems to be a key issue worth investigating to improve the current state of chatbots in the mental health domain.

PRIESTESS : Goedemiddag ik ben PRIESTESS, een vrouwelijke chatbot. Hoe heet je?
Good afternoon, I'm PRIESTESS, a female chatbot. What's your name?

User : eduard
eduard

PRIESTESS : Fijn dat je mee wil doen met dit onderzoek, Eduard!
Nice of you that you want to participate in this research, Eduard!

PRIESTESS : Ben je hiervoor al eens naar Lowlands geweest?
Have you been to Lowlands before?

User : ja
yes

PRIESTESS : Ah, tof dat je hier eerder bent geweest! Hoe is het dit jaar op Lowlands?
Ah, cool that you've been here before! How is Lowlands this year?

User : leuk!
fun!

PRIESTESS : Leuk om te horen dat deze Lowlands tof is! Dat komt ook door dit onderzoek natuurlijk. :heart:
Nice to hear that this year's Lowlands is cool! That's also because of this research of course :heart:

PRIESTESS : Welke optredens heb je tot nu toe gezien op Lowlands dit jaar?
Which shows have you seen at Lowlands this year?

User : ronnie flex onder andere, maar ook bijvoorbeeld boy pablo en franz ferdinand
ronnie flex amongst others, but also boy pablo and franz ferdinand for example

PRIESTESS : Ah, Boy Pablo! Je hebt dus wel interesse in bedroom pop. Hoe was het optreden van Boy Pablo?
Ah, Boy Pablo! So you're interested in bedroom pop. How was Boy Pablo's show?

User : leuk!
fun!

PRIESTESS : Klinkt goed! Ik ben blij dat het leuk was. :)
Sounds good! I'm happy it was fun. :)

PRIESTESS : Hé Eduard, ik vond het leuk om je beter te leren kennen. Maar je bent hier natuurlijk gekomen om te biechten. Vertel eens wat je wil opbiechten.
Hey Eduard, I liked getting to know you better. But you've come here to confess of course. Tell me what you want to confess.

User : ik heb de chocola van mijn huisgenoot gestolen
i stole my roommate's chocolate

PRIESTESS : Je maakt me hongerig met dat gepraat over eten. Wat voor gevoel heb je bij wat je gedeeld hebt?
You're making me hungry with that talk about food. What do you feel about what you've shared with me?

User : ik voel me schuldig
i feel guilty

PRIESTESS : Ai, vervelend dat je negatieve emotie ervaart. Wil je verder nog iets kwijt erover?
Ah, pity you're experiencing negative emotion. Do you want to share anything else about it?

User : nee
no

PRIESTESS : Dankjewel dat je wilde biechten!
Thanks for your confession!

Table 1. A (fictional) example of a conversation with PRIESTESS.

4 PRIESTESS: A chatbot for self-disclosure

4.1 System overview

As has been argued in the previous sections, more empirical evaluation research regarding chatbots in the mental healthcare domain is important in order to improve the quality of current e-mental health applications. Especially research with a focus on self-disclosure. Since most of the existing applications have not been shared publicly [3], we have developed **P**rogram **R**esponding to and **I**nquiring about **E**vents and **S**ecrets **T**old for the **E**valuation of **S**ecret **S**haring (hereafter: PRIESTESS). The goal of the system is to encourage users to share a secret that they have. The agent can be used on popular chat platforms *RocketChat* and *Discord* (both GDPR compliant). The system produces Dutch output, and uses emotion classification components based on LIWC [23] and Pattern [27] to interpret user input and provide feedback on secrets that users have shared. Furthermore, the chatbot is programmed to ask icebreaker questions beforehand to establish a relationship. These icebreaker questions serve to demonstrate the chatbot's level of empathy and understanding, which could facilitate self-disclosure. An example of a PRIESTESS conversation and its translation can be found in Table 1.

4.2 System architecture

Module	Explanation
Answer Understanding Module	Contains all the natural language understanding functions to interpret user's answers.
Governing Module	Goes over conversation topics, sends the messages linked to the topics, and activates relevant modules after receiving an answer.
Lookup Module	Opens the template database and retrieves all the questions and answers used by the chatbot.
Participant Number Module	Ensures every user gets a unique number assigned to them.
Template Filler Module.py	Finds and fills slots in the templates.
Template Filler Rule Module	Returns the relevant information to fill template slots.
Chat Interaction Module	Checks if a user has answered a question.
Extra Commands Module	Extra commands to clear the chat after a participant has finished talking to PRIESTESS, find the last message in the chat, and find the first response to PRIESTESS's questions.
Save Chat Module	Saves the chat.

Table 2. PRIESTESS's modules and descriptions of the modules

PRIESTESS's architecture is rule- and template-based. While the popularity of chatbots with a self-learning, neural architecture is on the rise [17], this more traditional approach does not require large amounts of training data and ensures that PRIESTESS's responses are consistent and controlled for. Such control is valuable when discussing sensitive topics, as can be expected in self-disclosure conversations. PRIESTESS's architecture is based on PASS, a system that generates soccer reports [18]. Like PASS, PRIESTESS is open-source and its modular structure makes it relatively easy to create extensions and adjustments.[1] Adjusting the content or translating the bot to another language is therefore a feasible task. An overview of these modules can be found in Table 2.

The system starts with the **Governing Module**, which takes all the discussion topics from the database (provided by the **Lookup Module**) and goes over them one-by-one. Each topic represents a question and a response to the user's answer. Thus, in the Governing Module, the chatbot starts each topic by sending a question to the user.

When the user answers this question, the **Answer Understanding Module** is initiated. This module contains all the natural language understanding functions that PRIESTESS applies to interpret the user's input and recognize the core information in this input. The interpretation of the Answer understanding Module is then returned to the Governing Module, which uses this information to find a relevant response to the user in the database.

Some of the questions and answers have the form of a template: a sentence with empty slots that can be filled with information. When a template is chosen, the **Template Filler Module** is activated, which finds these empty slots and sends them to the **Template Filler Rule Module**. The Template Filler Rule Module finds the relevant information for these slots and returns this to the Template Filler Module. This information is then used by the Template Filler Module to fill the empty slots. The filled template is finally given to the Governing Module to send a response to the user.

Other than that, there are some less impactful modules: the **Participant Number Module** ensures that every user gets a new and unique number assigned to them when the conversations are saved. For RocketChat specifically, there are separate modules to find user responses, and save the chat (the **Extra Commands Module** and **Save Chat Module**), as these functions require more effort for the RocketChat platform compared to the Discord platform.

5 Conclusion

We have provided an overview of the current state of mental health chatbots and its evaluation. Furthermore, we have argued the need for more research on self-disclosure as this may provide information about the role that chatbots should fulfill in the mental health landscape and society as a whole. An open-source chatbot was introduced that can be used for experiments on self-disclosure. We hope that this conversational agent may help facilitate research that can contribute to mapping the preferred role and content that a mental health chatbot should be using according to its users.

---

[1] `https://github.com/TallChris91/PRIESTESS`

Acknowledgements

We received support from RAAK-PRO SIA (2014-01-51PRO) and The Netherlands Organization for Scientific Research (NWO 360-89-050), which is gratefully acknowledged. We would also like to thank the anonymous reviewers for their valuable comments.

References

1. Aharony, N.: Relationships among attachment theory, social capital perspective, personality characteristics, and Facebook self-disclosure. Aslib Journal of Information Management **68**(3), 362–386 (2016)
2. Andersson, G., Cuijpers, P.: Internet-based and other computerized psychological treatments for adult depression: A meta-analysis. Cognitive Behaviour Therapy **38**(4), 196–205 (2009)
3. Arnberg, F.K., Linton, S.J., Hultcrantz, M., Heintz, E., Jonsson, U.: Internet-delivered psychological treatments for mood and anxiety disorders: A systematic review of their efficacy, safety, and cost-effectiveness. PloS one **9**(5), 1–13 (2014)
4. Barak, A., Hen, L., Boniel-Nissim, M., Shapira, N.: A comprehensive review and a meta-analysis of the effectiveness of internet-based psychotherapeutic interventions. Journal of Technology in Human Services **26**(2-4), 109–160 (2008)
5. Bell, S., Wood, C., Sarkar, A.: Perceptions of chatbots in therapy. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. ACM (2019)
6. 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 pp. 1–13 (2019)
7. Bird, T., Mansell, W., Wright, J., Gaffney, H., Tai, S.: Manage your life online: A web-based randomized controlled trial evaluating the effectiveness of a problem-solving intervention in a student sample. Behavioural and Cognitive Psychotherapy **46**(5), 570–582 (2018)
8. Chang, C.W., Heo, J.: Visiting theories that predict college students' self-disclosure on Facebook. Computers in Human Behavior **30**, 79–86 (2014)
9. Choi, Y.H., Bazarova, N.N.: Self-disclosure characteristics and motivations in social media: Extending the functional model to multiple social network sites. Human Communication Research **41**(4), 480–500 (2015)
10. Christensen, H., Griffiths, K., Evans, K., et al.: e-Mental health in Australia: Implications of the Internet and related technologies for policy. Commonwealth Department of Health and Ageing Canberra (2002)
11. D'Alfonso, S., Santesteban-Echarri, O., Rice, S., Wadley, G., Lederman, R., Miles, C., Gleeson, J., Alvarez-Jimenez, M.: Artificial intelligence-assisted online social therapy for youth mental health. Frontiers in Psychology **8**, 1–13 (2017)
12. Fitzpatrick, K.K., 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 Mental Health **4**(2), 1–11 (2017)
13. 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 Mental Health **5**(4), 1–15 (2018)

14. GGZ Nederland: Arbeidsmarktagenda GGZ - GGZ Nederland, maart 2018 (2018), https://www.ggznederland.nl/uploads/assets/Arbeidsmarktagenda%20GGZ%20Nederland%20-%20maart%202018.pdf
15. Ho, A., Hancock, J., Miner, A.S.: Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. Journal of Communication **68**(4), 712–733 (2018)
16. Lambert, M.J., Barley, D.E.: Research summary on the therapeutic relationship and psychotherapy outcome. Psychotherapy: Theory, Research, Practice, Training **38**(4), 357–361 (2001)
17. Laranjo, L., Dunn, A.G., Tong, H.L., Kocaballi, A.B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A.Y., et al.: Conversational agents in healthcare: a systematic review. Journal of the American Medical Informatics Association **25**(9), 1248–1258 (2018)
18. van der Lee, C., Krahmer, E., Wubben, S.: PASS: A Dutch data-to-text system for soccer, targeted towards specific audiences. In: Proceedings of the 10th International Conference on Natural Language Generation. pp. 95–104 (2017)
19. Liu, D., Li, Y., Thomas, M.A.: A roadmap for natural language processing research in information systems. In: Proceedings of the 50th Hawaii International Conference on System Sciences (2017)
20. Lucas, G.M., Gratch, J., King, A., Morency, L.P.: It's only a computer: Virtual humans increase willingness to disclose. Computers in Human Behavior **37**, 94–100 (2014)
21. Martínez-Miranda, J.: Embodied conversational agents for the detection and prevention of suicidal behaviour: Current applications and open challenges. Journal of Medical Systems **41**(9), 135–149 (2017)
22. Mou, Y., Xu, K.: The media inequality: Comparing the initial human-human and human-AI social interactions. Computers in Human Behavior **72**, 432–440 (2017)
23. Pennebaker, J.W., Booth, R.J., Francis, M.E.: Linguistic Inquiry and Word Count: LIWC2015. http://www.liwc.net Pennebaker Conglomerates, Austin, TX (2015)
24. Savin-Baden, M., Tombs, G., Bhakta, R.: Beyond robotic wastelands of time: Abandoned pedagogical agents and new pedalled pedagogies. E-learning and Digital Media **12**(3-4), 295–314 (2015)
25. Shum, H.Y., He, X.d., Li, D.: From Eliza to XiaoIce: Challenges and opportunities with social chatbots. Frontiers of Information Technology & Electronic Engineering **19**(1), 10–26 (2018)
26. Skjuve, M., Brandtzæg, P.B.: Chatbots as a new user interface for providing health information to young people. In: Andersson, Y., Dahlquist, U., Ohlsson, J. (eds.) Youth and news in a digital media environment – Nordic-Baltic perspectives, chap. 6, pp. 59–66. Nordicom, Göteborg, Sweden (2018)
27. de Smedt, T., Daelemans, W.: Pattern for Python. Journal of Machine Learning Research **13**(Jun), 2063–2067 (2012)
28. Taylor, D.A., Altman, I.: Self-disclosure as a function of reward-cost outcomes. Sociometry pp. 18–31 (1975)
29. Vaidyam, A.N., Wisniewski, H., Halamka, J.D., Kashavan, M.S., Torous, J.B.: Chatbots and conversational agents in mental health: A review of the psychiatric landscape. The Canadian Journal of Psychiatry **64**(7), 456–464 (2019)
30. Veerbeek, M., Knispel, A., Nuijen, J.: GGZ in tabellen 2013-2014. Trimbos-instituut (2015), https://www.trimbos.nl/docs/cfe358cd-59cc-4033-ac68-50711f835768.pdf