Position paper presented at CONVERSATIONS 2022 – the 6th International Workshop on Chatbot Research, Applications and Design, hosted by the University of Amsterdam, Netherlands, November 22-23.

Introducing a Framework for Designing and Evaluating Interactions with Conversational Agents

Mart Kicken[1], Chris van der Lee[2], Kim Tenfelde[2], Barbara Maat[1] and Jan de Wit[2]

[1] Elisabeth-TweeSteden Ziekenhuis, Department of Pharmacy
[2] Tilburg University, Tilburg School of Humanities and Digital Sciences, Department of Communication and Cognition
{m.kicken, b.maat}@etz.nl,
{c.vdrlee, k.tenfelde, j.m.s.dewit}@uvt.nl

Abstract. With chatbots and other types of conversational agents becoming more ubiquitous in everyday life, we see a need for tools and frameworks to structure the design and evaluation of this new technology, to optimize its effectiveness and user experience. From existing literature, we have identified eight domains from which the quality of conversational agents can be assessed: personality and usability, onboarding, understanding and context awareness, answering and response accuracy, navigation and engagement, error management, intelligence, and compatibility and information safety. We propose concrete tools and example questions that can be used to inform the design of chatbots, and to conduct expert evaluations and user studies. The framework is published as a living document in our supplementary materials.

Keywords: Chatbot design, Chatbot evaluation, Framework

1 Introduction

Recent advances in Natural Language Processing (NLP) and Artificial Intelligence (AI) transformed various aspects of our lives by utilizing enormous amounts of data and information to improve decision-making through problem solving, reasoning, and learning. As a result of these developments, the chatbot made its entry into diverse fields such as customer service, gaming, security, entertainment, travel, and healthcare in the last two decades [1–2]. Chatbots are computer programs designed to simulate human conversations with their users [3]. Their application is found useful in these multitudes of different domains, due to the ability for complex dialog management and conversational flexibility. Still, the rapid emergence and widespread use of this new technology calls for guidelines and frameworks. In this paper we present a framework based on a survey of existing literature, to ensure that these chatbots are effective and provide a pleasant user experience.

Chatbots are able to interact in a human-like fashion through four essential stages: input processing, input understanding, response generation and response selection. The user gives an input, which is received and processed by the chatbot followed by a selected output [1]. Furthermore, chatbots are either rule-based or use a subset of AI called machine learning (ML), which can be classified as retrieval or generative based chatbots [1,4,5]. *Rule-based chatbots* base their response on a script of predefined answers and rules that determine which answer to send. The first effective chatbot ELIZA, developed in 1966, was rule-based [2]. Responses of rule-based chatbots are hand-coded and fixed; they cannot answer any questions outside these predefined conditions. Hence, rule-based chatbots are rigid, difficult to scale, untrainable and require continuous (and most often expensive and intensive) maintenance [5]. This is in contrast with ML generating responses that are based on analyzing user input combined with a baseline script. Such chatbots provide a continuous stream of data in the interactions between the user and the chatbot. This data is given to a generic algorithm, improving itself with each interaction [1]. Regarding *retrieval-based* models, a chatbot is trained to provide ('retrieve') the best possible response from a database of predefined responses. Thus, retrieval-based models do not generate any new output and are limited to their database of response patterns [6]. A *generative chatbot* is the most sophisticated of all models. It trains itself using previous and current messages of users and generates original combinations of language based on probability distributions. However, it needs a very large dataset of user interactions and is prone to training bias [6–7]. Next to the method used to generate responses, chatbots can be further classified based on their primary goal, with chatbots either being *informative*, *conversational/chat-based,* or *task-based* [5].

Although chatbots are becoming ubiquitous, a clear and comprehensive framework for their design and validation is proven to be difficult and frequently specific for a particular chatbot's functionality or limited by the available technology. Furthermore, sample sizes of existing evaluation studies are often small, focusing on different outcome measures, and they are inconsistent in reporting characteristics of chatbots and study methods [2,8]. With increases in digitalization and the development of more and more chatbots, a more robust and standard framework used to perform expert, preliminary evaluations as well as user studies would be welcomed and most helpful.

To facilitate this need for structure, we propose validating chatbots based on eight different domains, which are explained in the section below. The different domains for testing are based on the consensus between literature studies [9], expert opinion [10], patients' perception [11] and leading AI chatbot testing tools (Botium [12]; Botanalytics [13]; Chatbottest [14]). Moreover, by combining findings from literature with practical applications of chatbot testing tools we aim to provide a comprehensive overview. The framework's purpose is to provide tools that can be applied more directly to the purpose and goals of an individual chatbot. Table 1 provides an overview of each domain, its description and purpose, and examples of useful ways of testing. A detailed version, as a living document, is available as supplementary materials[1]. Additionally, example questions are provided that can be used for chatbot evaluation, with the aim to contribute to standardizing evaluation practices in the field. By making it a living document, our framework will be continuously updated to any progress in the field.

By combining different existing frameworks our framework applies to either rule-based, retrieval, generative chatbots and to informative, conversational and/or task-based chatbots. Our framework differs from other frameworks in that it strongly defines

[1] https://osf.io/p2uve/

domains. For example, various studies have found that in the Natural Language Generation domain (which includes chatbots) there is very little agreement about which constructs are measured, and how they should be defined [15–16]. In addition, by focusing on conventions between different existing frameworks a more complete overview is formed.

2 Frameworks for testing chatbots

Multiple frameworks originating from different perspectives of relevant stakeholders were analyzed to formulate a general consensus in testing chatbots.

Firstly, existing literature and studies of chatbot testing were examined. Abd-Alrazaq *et al.* (2020) [9] performed an extensive review of 27 technical metrics used to evaluate chatbots, placed under four categories: general performance (e.g. usability, speed, intelligence, dialogue handling, error management), response generation (e.g. comprehensibility, realism, repetitiveness), response understanding (e.g. chatbot understanding as assessed by users, word error rate, concept error rate), and aesthetics (e.g. appearance of the virtual agent, background color, content).

Secondly, the identification of a consensus view across domain experts regarding testing of chatbots was considered. Denecke *et al.* [10] used a Delphi study design to find a consensus between professionals regarding relevant metrics to evaluate chatbots. Twenty-four metrics achieved high consensus and three metrics achieved moderate consensus, which were divided into four categories: global metrics (e.g. ease of use, security content accuracy), response generation (e.g. appropriateness of responses, realism, comprehensibility), response understanding (e.g. understanding, word error rate, attention estimation errors) and with less consensus aesthetics (e.g. font type and size, color).

Thirdly, perceptions and opinions of patients as end-users were analyzed. Abd-Alrazaq (2021) [11] performed an extensive review evaluating 37 unique studies and identified ten themes: usefulness, ease of use, responsiveness, understandability, acceptability, attractiveness, trustworthiness, enjoyability, content, and comparisons.

Lastly, three leading AI chatbot testing tools were investigated. Botium [12] is an eminent company specialized in testing conversational AI's. They performed more than 100 million tests of different chatbots with 1532 companies using Botium. They test for five different domains: functionalities (e.g. accuracy, compatibility, context understanding, error management), chatbot platform security, chatbot personalization, learning ability and accessibility. More specifically, CHARM is a module part of Botium [17] testing for conversational flow looking at: understanding, response and context integration. Botanalytics [13] is a world leading company for conversational analytics tools based in San Francisco. It evaluates chatbots on features: personality, user experience, conversation flow (e.g. understanding, answering), error management, onboarding, compatibility and security. Chatbottest [14] is an open source guide for testing chatbots based on seven different domains: personality, onboarding, understanding, answering, navigation, error management, intelligence.

Combining the metrics, categories, themes, and domains from literature and practices in the field reviewed in this section has resulted in the eight domains we propose for comprehensively assessing the quality of an interaction with a conversational agent. Note that we draw mostly from surveys in the field of health care. Nevertheless, findings were generalized to all types of conversational agents as we believe these metrics and themes to be widely relevant and applicable.

3 Domains for Testing Chatbots' Validity

3.1 Personality and Usability [2,9,10,12–14,18–24]

The chatbot personality defines the user's experience. Most often, the user prefers its conversation with the chatbot to imitate a human interaction. Having a comprehensive personality makes the chatbot more relatable, believable, and relevant to the user. Furthermore, a detailed personality formulates a deeper understanding of the chatbot's end goal(s), and how it will communicate through choices of language, tone, humor, and style [25]. The most optimal personality of a chatbot depends on the user (e.g. age group, gender, literacy) and task. The first step is to identify potential users of the chatbot, and to find a personality that fits its target audience best. A good first impression, and thus the very first message sent by the chatbot, is important. For example, a well-placed greeting will immediately break the ice between a bot and its user and will foster their engagement.

Furthermore, usability is the extent to which the chatbot can be used by its users to achieve the specific goals set by your institution. It poses a more general question: why is your chatbot needed? What are its goals exactly? The answer would likely also include factors related to the intended user experience (UX) and will define the tone, design, and purpose of your chatbot. In addition, a usable chatbot is easy to use without issues.

Besides testing using private testing tools [12–14], there are multiple validated questionnaires available (see Table 1). The most well-known and widely-used usability questionnaire is the Systems Usability Scale (SUS) [9,26], although its applicability to chatbots has been disputed [27]. Lastly, Kocabalil *et al.* [18] provide a comprehensive overview of each questionnaire and its testing of different parts of UX.

3.2 Onboarding [9,10,13,14]

Onboarding is important in drawing and keeping the attention of the user throughout each interaction. It is dependent on the **clarity** and **presentation** of the chatbot. Regarding clarity, the chatbot should be easily accessible and use a clear way of communication. At all times and throughout each interaction, the user should know what is expected of them and know what they can expect from the chatbot. Additionally, the presentation of the chatbot should be simple, straightforward, and easily understandable. Important parts of its presentation include adequate profile description/picture, human-like interactions, and suitable appearances through its interactions and tone. Examples of testing for onboarding are the Chatbot Usability Questionnaire (CUQ) [27], AttrakDiff [20] and Subjective assessment of speech-system interface (SASSI) [21] in addition to Botanalytics [13] and Chatbottest [14].

3.3 Understanding and Context Awareness [9,10,12–14,17,28,29]

The understanding of the chatbot will determine its responses, whereas effective context awareness will provide the chatbot with the capability of interacting with users in an efficient, intelligent, and natural way. Different kinds of inputs will require different kinds of processing. For example, small mistakes in spelling by the user should still be understandable to the chatbot. However, the chatbot should not only focus on understanding one word, but also all the words which are near (linguistic context). This will create context and provides more relevant responses of the chatbot.

Examples of testing on understanding and context awareness are CUQ (questionnaire), CHARM and Botium [12,17], seq2seq [28,29] and Beam Search Decoding [29]. By coding each word into specific vectors and analyzing each sentence, these tests will identify the most possible and likely sequence as output. In addition, Botanalytics [13] and Chatbottest [14] also provide testing on these domains.

3.4 Answering and Response Accuracy [9,10,12–14,17,28,29]

Adequate responses of the chatbot to the user ensures engagement and a smooth conversation flow. The chatbot's output will differ and should be appropriate depending on the chatbot's purpose and use, as well as its target audience.

This domain is an extension of the previous domain, understanding and context awareness. By improving the chatbot's understanding of the input of the user, it can formulate a more accurate response and/or answer. Hence, there are some overlaps between (response) accuracy and the previous domain of understanding and context awareness. The main difference lies in how well the chatbot answers. Examples of testing are CUQ (questionnaire) [27] or Botanalytics [13] and Chatbottest [14].

3.5 Navigation and Engagement [10,13,14,20–23]

It is important to maintain a simple navigation flow, so that the user does not feel lost at any time in the process. At all times the user should be able to go back to previous questions or sections. In addition, navigating easily through different parts/flows of the chatbot is important for potential revision of previously answered questions.

Furthermore, engagement during interactions with the chatbot is crucial for these interactions to be effective. Engagement will make the user more involved and stimulates attention. For example, if the chatbot goes through a standard set of questions, transparency of the process (e.g. how many questions are left) is essential for the engagement of the user.

This domain is often linked to usability and user experience. To evaluate navigation and engagement throughout the conversation, determining the user's experience is therefore important. Possible useful questionnaires in descending extensiveness are AttrakDiff [20], SASSI [21], SUISQ [23] and MOS-X [22] and both testing tools Chatbottest [14] and Botanalytics [13].

3.6 Error Management [9,10,13,14]

Chatbots are not perfect. However, through using ML, chatbots have the potential to study errors to improve their interaction with users. Therefore, error management is of utmost importance to an effective and continuously learning chatbot.

Error management provides insights into the way the chatbot responds and performs when something unexpected happens. It provides an indication of the extent to which the chatbot can adapt after, and if it can provide a clear analysis of the error or breakdown in communication. Error messages therefore should describe the error to the user. Rather than repeating the same messages or question, chatbot error handling should clarify the error, and propose a way to move forward. For example, if the chatbot is confused by out-of-context answers of the user, the chatbot could remind the user of its capabilities in responses. Moreover, differences between internal and external errors should be made.

Unfortunately, literature for testing error management is scarce or only limited to observations and evaluations of numbers of errors [9,10]. An example is CUQ [27] that poses only one question ("The chatbot coped well with any errors or mistakes"). However, both Botanalytics [13] and Chatbottest [14] extensively test for error management.

3.7 Intelligence [9,12–14,30]

Intelligence represents the possibility of the chatbot to analyze interactions with users and to be able to learn from them. The chatbot recognizes relevant patterns and correlations and uses data provided by its interactions to improve itself. Important to note here is that an intelligent chatbot will be only as good as the data it was trained with. For example, there could be a risk of chatbots being 'reprogrammed' by users due to the chatbot being incomplete and therefore receiving incorrect information of users.

Testing should be based on users' experiences and identifying the possibility of the chatbot to learn from its interactions. Regarding the perception of the chatbot by the user, questionnaires such as CUQ [27] and Bartneck *et al.*'s [30] questionnaire called Godspeed (that includes various factors such as perceived intelligence and anthropomorphism), can be used. Furthermore, Botium [12], Botanalytics [13] and Chatbottest [14] all explicitly test for intelligence.

3.8 Compatibility and safety [10,12–14]

Due to many different software and hardware options nowadays, it is important that the chatbot will be compatible with each one of them. Software of the chatbot should run on different platforms (e.g. smartphone, tablet, laptop, desktop) and on different operating systems (Windows, OS X, Android, iOS).

Moreover, safety of personal information of users and cyber security should always be of high concern. There should be a continuity in ensuring safe (personal) data storage. Data should be encrypted to be anonymous and untraceable to the user. Additionally, the chatbot should not provide any sensitive information of other users to a user. Literature for testing compatibility and safety in chatbots is limited. For now, only private chatbot testing is available (Botium [12], Botanalytics [13] and Chatbottest [14]).

4 Conclusion and Recommendations

The last couple of years the usage of chatbots in various domains has increased. Chatbots are used more and more in supporting and relieving tasks previously performed only by people. However, there was not yet a robust or standardized framework for validating chatbots' effectiveness. Based on consensus of researchers, domain experts, end-users and leading AI chatbot testing tools Botium, Botanalytics and Chatbottest, we proposed a framework for validating chatbots based on eight different domains: personality and usability, onboarding, understanding and context awareness, answering and response accuracy, navigation and engagement, error management, intelligence, and compatibility and safety.

Most of these domains can be validated using questionnaires for user experience, however some of them rely on testing tools only. Hence, more research and validated checklists and questionnaires are needed to support and facilitate validating chatbots on these domains. With this position paper we propose a framework and tools to apply specifically to the purpose and goals of an individual chatbot, and pave the way forward for new tools to be developed. We recommend usage of a framework for validating chatbots based on these eight different domains, which can also be used as a guideline to formulate a clear definition of chatbot requirements. It is important to note that the domains share some overlap, and that it may not be necessary for every conversational agent to meet all of the proposed guidelines fully.

A combination of different questionnaires can be used to quantify and qualify user experiences throughout all domains except error management and compatibility and safety. However, these domains can be tested using chatbot testing tools. For this we recommend the usage of Chatbottest as an open source, free to use testing tool. The concrete example questions we pose in the full version of the table can also be used in an expert evaluation of a chatbot under development.

Since literature, especially in the field of chatbot testing, is continuously updated, we introduce a living document. Our framework is easily adjusted to any progress in the field of chatbot testing. Hence, this position paper is meant to be a first draft and as a measurement of potential interest, making it hopefully a first step for long term development of robust chatbot testing tools.

Lastly, throughout these domains the adage always goes: "Simplicity is the ultimate sophistication." It is better to make your chatbot as easily useable and simple as possible!

Table 1. Domains for testing chatbots validity. The full table is available as a 'living document': https://osf.io/p2uve/

Domains	Description	Tools and questionnaires
1. Personality and usability	Does the chatbot have a clear voice and tone that fits with the users and with the ongoing conversation?	CUQ [27]; SUS [9,26]; AttrakDiff [20]; SASSI [21]; SUISQ [23]; MOS-X [22]; PARADISE [24]; Botium [12]; Botanalytics [13]; Chatbottest [14]
2. Onboarding	Is it clear for users from the very beginning what the chatbot is about? Does the chatbot present itself in an open and simple manner?	CUQ [27]; AttrakDiff [20]; SASSI [21]; Botanalytics [13]; Chatbottest [14]
3. Understanding and context awareness	How is the understanding of the chatbot of different inputs from the user? Is the chatbot aware of different contexts?	CUQ [27]; CHARM [17]; Botium [12]; Seq2seq [28,29]; Beam Search Decoding [29]; Botanalytics [13]; Chatbottest [14]
4. Answering and response accuracy	What kind of answers and responses does the chatbot give and how relevant are these? Are they relevant to the moment and context?	CUQ [27]; Botanalytics [13]; Chatbottest [14]
5. Navigation and engagement	How easy is it to operate through the conversation and its different flows/parts? Does the user remain engaged throughout the process?	AttrakDiff [20]; SASSI [21]; SUISQ [23]; MOS-X [22]; Botanalytics [13]; Chatbottest [14]
6. Error management	How does the chatbot deal with errors that occur? Is it able to recover from them?	CUQ [27]; Botium [12]; Botanalytics [13]; Chatbottest [14]
7. Intelligence	Does the chatbot have any intelligence? Would it pass a Turing test?	CUQ [27]; Godspeed [30]; Botium [12]; Botanalytics [13]; Chatbottest [14]
8. Compatibility and information safety	Can the chatbot be used on different devices and is (personal) data stored safely?	Botium [12]; Botanalytics [13]; Chatbottest [14]

References

1. Xu, L., Sanders, L., Li, K. & Chow, J. C. L.: Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review. *JMIR Cancer 7*(4), e27850 (2021).
2. Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S. & Coiera, E.: Conversational agents in healthcare: a systematic review. *J. Am. Med. Inform. Assoc. 25*(9), 1248–1258 (2018).
3. Dahiya, M.: A Tool of Conversation: Chatbot. *Int. J. Comput. Sci. Eng. 5* 158–161 (2017).
4. Parviainen, J. & Rantala, J.: Chatbot breakthrough in the 2020s? An ethical reflection on the trend of automated consultations in health care. *Med. Health Care Philos. 25*(1), 61–71 (2022).
5. Adamopoulou, E. & Moussiades, L.: An Overview of Chatbot Technology. in *Artificial Intelligence Applications and Innovations* (eds. Maglogiannis, I., Iliadis, L. & Pimenidis, E.) 373–383 (Springer International Publishing, 2020). doi:10.1007/978-3-030-49186-4_31.
6. Hien, H. T., Cuong, P.-N., Nam, L. N. H., Nhung, H. L. T. K. & Thang, L. D.: Intelligent Assistants in Higher-Education Environments: The FIT-EBot, a Chatbot for Administrative and Learning Support. in *Proceedings of the Ninth International Symposium on Information and Communication Technology* 69–76 (Association for Computing Machinery, 2018). doi:10.1145/3287921.3287937.
7. Patidar, M., Agarwal, P., Vig, L. & Shro, G.: Correcting Linguistic Training Bias in an FAQ-bot using LSTM-VAE. 16.
8. Abd-alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P. & Househ, M.: An overview of the features of chatbots in mental health: A scoping review. *Int. J. Med. Inf. 132* 103978 (2019).
9. Abd-Alrazaq, A., Safi, Z., Alajlani, M., Warren, J., Househ, M. & Denecke, K.: Technical Metrics Used to Evaluate Health Care Chatbots: Scoping Review. *J. Med. Internet Res. 22*(6), e18301 (2020).
10. Denecke, K., Abd-Alrazaq, A., Househ, M. & Warren, J.: Evaluation Metrics for Health Chatbots: A Delphi Study. *Methods Inf. Med. 60*(5/6), 171–179 (2021).
11. Abd-Alrazaq, A. A., Alajlani, M., Ali, N., Denecke, K., Bewick, B. M. & Househ, M.: Perceptions and Opinions of Patients About Mental Health Chatbots: Scoping Review. *J. Med. Internet Res. 23*(1), e17828 (2021).
12. Botium Box - The chatbot testing tool. *Botium* https://www.botium.ai/.
13. Botanalytics | AI powered Chatbot Analytics and Voice Analytics. https://botanalytics.co.
14. Home · chatbottest-com/guide Wiki. *GitHub* https://github.com/chatbottest-com/guide.
15. van der Lee, C., Gatt, A., van Miltenburg, E. & Krahmer, E.: Human evaluation of automatically generated text: Current trends and best practice guidelines. *Comput. Speech Lang. 67* 101151 (2021).
16. Howcroft, D. M., Belz, A., Clinciu, M.-A., Gkatzia, D., Hasan, S. A., Mahamood, S., Mille, S., van Miltenburg, E., Santhanam, S. & Rieser, V.: Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions. in *Proceedings of the 13th International Conference on Natural Language Generation* 169–182 (Association for Computational Linguistics, 2020).
17. Bravo-Santos, S., Guerra, E. & de Lara, J.: Testing Chatbots with Charm. in *Quality of Information and Communications Technology* (eds. Shepperd, M., Brito e Abreu, F., Rodrigues da Silva, A. & Pérez-Castillo, R.) 426–438 (Springer International Publishing, 2020). doi:10.1007/978-3-030-58793-2_34.
18. Kocabalil, A. B., Laranjo, L. & Coiera, E.: Measuring User Experience in Conversational Interfaces: A Comparison of Six Questionnaires. in (2018). doi:10.14236/ewic/HCI2018.21.
19. Larbi, D., Denecke, K. & Gabarron, E.: Usability Testing of a Social Media Chatbot for Increasing Physical Activity Behavior. *J. Pers. Med. 12*(5), 828 (2022).

20. Hassenzahl, M., Burmester, M. & Koller, F.: AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. in *Mensch & Computer 2003: Interaktion in Bewegung* (eds. Szwillus, G. & Ziegler, J.) 187–196 (Vieweg+Teubner Verlag, 2003). doi:10.1007/978-3-322-80058-9_19.
21. Hone, K. S. & Graham, R.: Towards a tool for the Subjective Assessment of Speech System Interfaces (SASSI). *Nat. Lang. Eng. 6*(3–4), 287–303 (2000).
22. Polkosky, M. D. & Lewis, J. R.: Expanding the MOS: Development and Psychometric Evaluation of the MOS-R and MOS-X. *Int. J. Speech Technol. 6*(2), 161–182 (2003).
23. Polkosky, M. D.: Machines as mediators: The challenge of technology for interpersonal communication theory and research. in *Mediated Interpersonal Communication* (Routledge, 2008).
24. Hone, K. & Graham, R.: Subjective assessment of speech-system interface usability. *Seventh Eur. Conf. Speech Commun. Technol.* (2001) doi:10.21437/Eurospeech.2001-491.
25. McTear, M., Callejas, Z. & Griol, D.: Emotion, Affect, and Personality. in *The Conversational Interface: Talking to Smart Devices* (eds. McTear, M., Callejas, Z. & Griol, D.) 309–327 (Springer International Publishing, 2016). doi:10.1007/978-3-319-32967-3_14.
26. Brooke, J.: SUS - A quick and dirty usability scale. *Usability Eval. Ind. 189*(194), 4–7.
27. Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V. & Mctear, M.: Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? in *Proceedings of the 31st European Conference on Cognitive Ergonomics* 207–214 (ACM, 2019). doi:10.1145/3335082.3335094.
28. Bozic, J., Tazl, O. A. & Wotawa, F.: Chatbot Testing Using AI Planning. in *2019 IEEE International Conference On Artificial Intelligence Testing (AITest)* 37–44 (2019). doi:10.1109/AITest.2019.00-10.
29. Reddy Karri, S. P. & Santhosh Kumar, B.: Deep Learning Techniques for Implementation of Chatbots. in *2020 International Conference on Computer Communication and Informatics (ICCCI)* 1–5 (2020). doi:10.1109/ICCCI48352.2020.9104143.
30. Bartneck, C., E, C. & D, K.: Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. (2008) doi:10.6084/m9.figshare.5154805.