doi: 10.36689/uhk/hed/2021-01-058 HED Hradec Economic Days

Measuring Chatbot Effectiveness

Hana MOHELSKA and Marcela SOKOLOVA*

University of Hradec Králové, Hradec Králové, Czech Republic; hana.mohelska@uhk.cz; marcela.sokolova@uhk.cz

* Corresponding author: marcela.sokolova@uhk.cz

Abstract: The paper deals with a technology called chat robot. Chatbot is one of the Fourth Industrial Revolution trends, and so far, there is no commonly used approach that has proven to measure its effectiveness. The aim of this paper is to delve deeper into this issue, describe the current state, and further analyse possible known alternatives for measuring the chatbot effectiveness and identify a suitable approach according to selected criteria. The translated paper's main aim is to select a suitable tool for measuring the effectiveness of chatbot. A literature search was carried out within the paper's elaboration, quantitative and qualitative research was gradually used to select a suitable tool for measuring the effectiveness of the chatbot, while specific data collection techniques are a document study, questionnaire survey in the form of pairwise comparison, where output is an expert opinion and semi-structured interview. Furthermore, a multi-criteria analysis is performed in the Expert Choice 2000 decision support tool. User friendliness, Information ability and equipment, Language level and equipment, Humanity, Business aspect were chosen as the most important criteria. The 2018 approach by the Croatian author D. Peras was chosen to carry out an analysis of approaches to measuring chatbot effectiveness.

Keywords: chatbot; efficiency; measurement; approach

JEL Classification: M14; M31; O31

1. Introduction

The advent and adoption of the "Industry 4.0" initiative, or simply the fourth industrial revolution, has gradually changed and continues to change the thinking philosophy of today's companies. Under this concept, one can imagine the transformation of production into a fully integrated automated and continuously optimised environment, which regards the digitisation of industrial production, but a comprehensive system of changes associated with phenomena such as the Internet of Things, Services and People, autonomous robots and artificial intelligence development, Big Data analysis, digital twin, virtualisation, cloud computing as well as augmented reality. The third industrial revolution gave companies the form of information, i.e., information knowledge, which consists in the ability to manage information efficiently, the fourth industrial revolution then follows and enables the use of newly introduced technologies to streamline the use of resources significantly.

This paper deals with a technology called chat robot, more commonly known as chatbot, which can be included among the above trends. Although its origin dates back to the 1960s, it generally began to find its use over the last few years, in various forms (Kotoučková, 2020),

(Bureš et al., 2012). Due to its abundant deployment, it is often discussed whether it really fulfils its purpose or it is only a utopian idea of the robot's ability to effectively replace human activity. In order for a company to be able to answer this question, it needs to have a well-established process for measuring this technology's effectiveness. However, a unified approach that would be commonly used has not yet been developed, and this is the issue the paper addresses.

2. Research Objective and Methodology

The translated paper's main aim is to select a suitable tool for measuring the chatbot effectiveness. The following research question was determined in line with this aim: ***RQ: "How is it possible to measure the effectiveness of a chatbot?"***

The following sub-goals must be met to fulfil the main aim:

1) Definition of the term chatbot - Literature Review
2) Determining the evaluation criteria of tools for measuring chatbot effectiveness
3) Identification of tools that can measure chatbot effectiveness
4) Selection of a suitable tool for measuring the effectiveness of the chatbot

A search of professional literature was carried out as part of this paper's elaboration, first focusing on the issue of chat robots, then on measuring the effectiveness of IS/ICT and approaches to measuring the chatbot effectiveness. Both printed and electronic sources were used, mainly foreign studies obtained from scientific databases, such as Science Direct, Springer, Web of Science. Their searches used the keywords such as "chatbot", "effectiveness", "measurement", "assessment", "evaluation", "metrics", in various combinations. Quantitative and qualitative research was gradually used to select a suitable tool for measuring chatbot effectiveness, with specific data collection techniques being document study, a questionnaire survey in the form of a pairwise comparison method, where the output is an expert opinion and a semi-structured interview. Furthermore, a multi-criteria analysis is performed in the Expert Choice 2000 decision support tool, whereby a suitable approach was selected to help measure chatbot effectiveness. Among other things, it is desirable to determine the weights of the individual criteria. For their objective determination, six interested experts were used who were subjected to a survey in the form of a pairwise comparison of defined criteria. Based on the individual results, Saaty matrices were compiled which led to the final output in the form of an expert opinion of the weights of the given criteria.

3. Literature Review

3.1. Chatbot - History and Classification

The term chatbot is sometimes also referred to as "chatterbot" or just "chat robot", and you can find several different definitions, which say that it is a technology automating or simulating human conversation or directly calling it artificial intelligence, as explained in the following subchapters, which is not entirely true. In conjunction with the above, Shawar and Atwell (2007) define this technology in an acceptable and clear manner as a computer programme that mediates interaction between it and a living person using natural language.

This means that it should not be recognised that a person is communicating with a robot during the process, and therefore there should be a completely immediate transfer of information between the two parties. Brandtzaeg and Følstad (2017) further state that the interaction occurs through chat interfaces, which companies often place on their websites or use existing platforms such as Facebook Messenger, WhatsApp, Skype, Slack and others. Some include voice-powered virtual assistants such as Siri, SVoice, Google Assistant, Cortana and Alexa (Rieke, 2018).

Although chatbot has become a trend in recent years (Accenture, 2016), Lokman and Zain (2010) state that its origins date back to 1966, when Professor Joseph Weizenbaum introduced the first chat robot, called ELIZA, whose behaviour was based on the simple principle of searching and creating outputs according to the keywords of the given inputs, subjected to the so-called decomposition (transformation) rules, and therefore later it became an inspiration for other chatbots. In 1995, there was a major breakthrough in this technology when Dr. Richard S. Wallace came up with his modern version of ELIZA, named A.L.I.C.E., whose name is derived from the Artificial Linguistic Internet Computer Entity. It then gave rise to a new Artificial Intelligence Markup Language (AIML), based on XML (eXtensible Markup Language) dialect, creating naturally speaking software agents. Furthermore, the authors mention that the idea of whether machines can think and how to verify this fact came in 1950, the mathematician Alan Turing, whose test consists in recognising whether the evaluator (real person) recognises whether they are conversing with a chatbot or a human individual. More precisely, everything takes place in two separate rooms, in one room there is an evaluator and in the other a human and artificial one. The interviewer asks questions to which they receive an answer. If a chatbot conversation was not detected, the machine passed the Turing test. Although A.L.I.C.E. did not pass this test, it won the Loebner Prize three times, in the years 2000, 2001 and 2004. The Loebner Prize has been an annual competition since 1991 based on passing the Turing test. Wakefield (2019) explains the rules of the game, whereby each machine is asked the same 20 questions with varying degrees of complexity, conversations last 25 minutes and the winner is the one who can convince more than half of the evaluators that they are human. However, so far, this has never succeeded, so the machine that managed to outsmart most of the evaluators has won the prize. Russel and Norvig (2010) mention strict criticism from Shieber (1994) regarding the usefulness of this test in the Loebner Prize competition. From ELIZA, through A.L.I.C.E. and other successful chatbots, such as the current five-time Loebner prize winner of the Mitsuku chatbot, created by Steve Worswick (Wakefield, 2019). In today's digital world, chatbots are becoming a common part of all types of companies.

3.2. Connection with Artificial Intelligence

The term Artificial Intelligence (hereinafter "AI") cannot be clearly defined. It is an interdisciplinary science "about the creation of machines or systems that will use a procedure in solving a certain task, which - if carried out by man - would be considered a manifestation of their intelligence" (Minsky, 1967). This definition is based on the Turing test and can be freely translated so that the complexity of solved tasks requires the use of human intelligence.

In this case, it is necessary to focus on the question of what the complexity and intelligent solutions are. Čermák (2018) states that complexity is characterised by the number of all possible solutions and the second attribute is limited by knowledge. He adds that artificial intelligence deals with the search for boundaries, including the representation of acquired knowledge and processes, the acquisition and use of it in solving problems, and uses various approaches and algorithms to find the basis of very complex tasks.

The basic question of whether machines can think had already been addressed in the 17th century by philosophers such as Descartes, Pascal, Hobbes, whose ideas only moved on a theoretical level and they were unable to turn it into reality. Furthermore, Čermák (2018) emphasises 1950, the year where Alan Turing came up with his test, whereby he gathered a number of pieces of evidence regarding intelligent machines, and subsequently refuted them. A conference led by John McCarthy was held in 1956 at Dartmouth College, bringing together experts interested in the possibility of implementing human thinking on machines (computers). The author mentions other important milestones in the history of artificial intelligence development, including A. Newell, H. Simon and their programme based on heuristic search techniques - the Logic Theorist, a system that mimics human thinking in solving problems in state space by reducing differences - General Problem Solver (GPS).

In 1958, John McCarthy created a language for artificial intelligence - LISP.

In the early 1970s, the PROLOG language was created by A. Colmerauer. Furthermore, various universal systems were designed during this period that were unable to solve highly specialised tasks. These include Planner, for example. The fundamental problem regarding artificial intelligence, namely the representation, use and processing of knowledge, began to be pointed out.

This led to the development of expert systems whose task is to simulate the decision-making process of specialists in solving complicated tasks using a knowledge base. Successful pioneers of expert systems included MYCIN and PROSPECTOR.

In 1981, the 5th Generation Computer Project for non-numerical information processing was announced in Japan. These computers worked with extensive knowledge bases and required learning mechanisms. Furthermore, T. Kohonen introduced the "electronic typewriter" in 1988. It involved the translation of voice into English text via cohonenon neural networks. In 1990, L. Adleman began to create the concept of computer DNA and demonstrated the subsequent feasibility of basic mathematical operations and calculations.

Shah (2017) states the issue of natural language processing is linked to chatbot and artificial intelligence which can be found under the abbreviation NLP (Natural Language Processing). Chatbots are able to function using predefined rules based on language structures (so-called rule-based) or using a statistical model of natural language processing, which deals with the so-called machine learning.

4. Chatbot Application Areas

Chatbot has largely found its application in the last few years, more precisely the biggest breakthrough of its widespread deployment came in 2016, when Facebook and Microsoft began to officially support the use of robots within their platforms (Khorozov, 2017). Business

Insider Intelligence (2016) presents the results of a survey conducted by Oracle, and according to which, in 2016 approximately 80% of surveyed American companies owned or planned to launch a chatbot by 2020 at the latest. Nicastro (2018) states that Inbenta, based on its survey of consumer and business chat robot usage and perception, rated a 50% preference for chatbot communication when shopping online over customer support calls, with 72% of shoppers finding its services to be error-free and very helpful. In the Czech Republic, the Feedyou agency is proud of several cases of successful chatbot implementations in the corporate environment, which since 2014 has been trying to show how this technology can benefit today's businesses. In areas such as HR, customer support, sales or even GDPR, it has helped companies such as ČEZ, a. s., STRV, s. r. o., Fincentrum, a. s., Knorr-Bremse, s. r. o. and others (Feedyou, 2019).

In order to be able to realistically consider the introduction of technology as successful or effectively fulfilling its purpose (not only because it is a trend), it is necessary to regularly measure its effectiveness. In this respect, a unified approach has not yet been developed, that companies could use to assess their chatbot, which is the subject of the following part of the paper. Other passages deal with the effectiveness of IS/ICT in general and further present in detail the approaches to measuring the effectiveness of chatbot, which are based on foreign studies dealing with this issue.

4.1. ICT/IS Effectiveness

Molnár (2000) states that it only makes sense to examine systems for which a purpose can be defined in terms of information system effectiveness (hereinafter "IS") and information technology (hereinafter "IT"), i.e., information and communication technologies (ICT). He then calls these systems as target behaviours. He also explains the relationship between information system and information technology, where information system represents the need for information, while information technology represents the satisfaction of this need and hence the abbreviation IS/IT, i.e., ICT. Furthermore, the author adds that the evaluation of efficiency does not only address the issue of needs and their fulfilment, but also the expectations of the involved parties. From a company-wide point of view, these can be:

- owners who see the introduction of IS/ICT as a permanent appreciation of the assets invested in the company,
- managers who are able to effectively manage the company on the basis of IS/ICT, without spending extra resources entrusted to them for administration,
- employees to whom IS/ICT offers benefits in the form of a better, more efficient and fully integrated work environment,
- customers to whom the use of IS/ICT brings higher added value of the required product or service.

It is important to note that the choice and combination of indicators always depends on the specific case, and to emphasise that the most relevant relationship for evaluating effectiveness is the monitoring of effectiveness aspect, which is directly measurable by the degree to which objectives are achieved, i.e., as follows:

Effectiveness = achieved goal value / planned goal value.

5. Determining the Criteria for Selecting a Suitable Tool for Measuring Chatbot Effectiveness and Designing Suitable Tools

To fulfil the paper's aim, it is necessary to choose a suitable approach for measuring chatbot effectiveness. It is selected using an analytical decision support tool called Expert Choice 2000. According to Expert Choice (2020), the first step in multi-criteria decision-making is to define the goal that should be achieved.

5.1. Determining the Criteria for Selecting a Suitable Tool for Measuring the Effectiveness of a Chatbot

Subsequently, it is necessary to establish selection criteria. This was selected on the basis of a careful study of professional studies dealing with the issue of chatbot evaluation and the advice of experts from ČSOB. The five most important areas are selected as criteria, which, according to the already mentioned theoretical background and expert determination, should not be missing in the right approach to measuring chatbot effectiveness. These include:

- user friendliness - evaluating the impression of the chatbot (incl. chat interface) on the user and whether they are satisfied with their services during the interaction,
- information ability and equipment - finding out and supplementing the state of the chatbot knowledge base so that the produced outputs can satisfy the user's needs,
- language level and equipment - evaluation of the chatbot's ability to create correct and verbally diverse outputs in terms of spelling and grammar,
- humanity - assessing whether the chatbot behaves like a human,
- business aspect - finding out what added value chatbot brings for a company and the metrics associated with it.

5.2. Determining Individual Criteria Weights

The established criteria can be considered important, but not to the same extent. Therefore, it is necessary to determine appropriate weights for each of them. For this, questionnaire survey results were used, in the form of a pairwise comparison method performed within the company. More precisely, the research involved a total of six selected managers and analysts, interested in this issue, and therefore capable of relevant assessment (they cover the administration, operation, support and all processes related to a chatbot). The survey's aim was to find out the importance that managers attach to the defined criteria, on the basis of which their final weights were determined, i.e., the expert opinion.

The questionnaire was created according to the Saaty method of pairwise comparison and the form of its processing was inspired by the Expert Choice 2000 programme itself. The data obtained from the performed research were processed into partial Saaty matrices. First, a matrix was created for each expert separately, whereby the output was the calculation of the partial and total geometric diameter and subsequent weights for the first to fifth criteria. These calculated weights were transferred to a common matrix showing their relationship from the first up to the sixth respondent. Then, an arithmetic mean was calculated from the

individual weights for each criterion, giving the values of the resulting (uniform) weights, which can be seen in Table 1.

Table 1. Determined weights of criteria and their final order. Source: own processing.

Criterion	Weight	Sequence
User friendliness	**0.214**	3.
Information ability and equipment	**0.345**	1.
Language level and equipment	**0.099**	4.
Humanity	**0.045**	5.
Business point of view	**0.297**	2.

As can be seen, of all five criteria, respondents consider the most important information ability and equipment - to satisfy the needs of existing or potential clients. Therefore, ensuring information capability and equipment represents added value for the customer and immediately behind it is the business aspect, which is formed by metrics evaluating the added value for the company. The weights of both criteria separately are around one third of the five criteria and therefore together represent a nearly two-thirds preference over them. Therefore, metrics focused on these two aspects should definitely not be missing in a suitable approach to measuring chatbot effectiveness.

User friendliness is seen as the third most important criterion, this requires an evaluation on how the chatbot acts on the user and how satisfied they are with its services. For this reason, it is necessary to focus on ensuring the possibility of gaining the most pleasant and effective experience that the user receives during the chatbot interaction.

In terms of importance, language and equipment are in the fourth place. This means assessing the chatbot's ability to produce correct and verbally diverse outputs from the viewpoint of spelling and grammar, thanks to which the user is able to capture an accurate interpretation of the transmitted information.

Humanity is considered to be the least important criterion. In the case of artificial intelligence, humanity testing plays an essential role, the chatbot is designed to be able to fulfil a specific task. The user already knows in advance that they are conversing via a chatbot and not a live operator, so the company pays more attention to other, in this case more important, criteria.

5.3. Choice of Alternatives

According to combinations of the keywords "chatbot", "effectiveness", "measurement", "assessment", "evaluation", "metrics", professional scientific databases such as Springer, Science Direct, Web of Science have found little the number of publications whose content would be usable for the purposes of the paper. A total of seven studies were selected by the method of analysis and subsequent synthesis, which dealt with at least two areas of measurement, i.e., selection criteria and included useful metrics. By discussing the most relevant findings, four were selected from seven studies, which offered the most comprehensive approach to measuring the chatbot effectiveness, i.e., dealt with at least four

of the five required measurement areas. The following alternatives were chosen to select a suitable approach to measuring chatbot effectiveness:

- Kuligowska (2015),
- Maroengsit et al. (2019),
- Peras (2018),
- Radziwill and Benton (2017).

6. Conclusions – Evaluating the Selection of an Appropriate Approach to Measuring Chatbot Effectiveness

Using the multi-criteria decision-making process implemented in the vAHP software product Expert Choice 2000, the following two graphs were created (see Figures 1 and 2) showing the suitability of selected approaches to measuring the effectiveness of the chatbot.

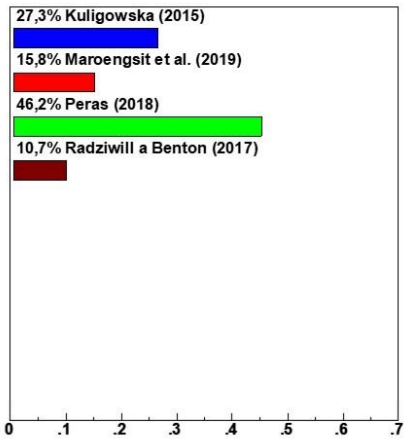

Figure 1. The result of selecting a suitable approach to measuring chatbot effectiveness in percentage terms. Source: own processing in the Expert Choice 2000 programme.

Figure 1 shows the differences in the suitability of individual approaches in percentage expression. Based on the multi-criteria analysis results, it is clear that the best rated approach is from Peras (2018), which achieves almost 50% fulfilment of the evaluated criteria compared to other alternatives. On the contrary, only 10% of the criteria were met by Radziwill and Benton (2017), which was described as insufficient and therefore unsatisfactory in terms of the three out of five criteria.

The most suitable approach out of the four analysed was the method of measurement created by the author Peras (2018), who divided the process of evaluating the effectiveness of chatbot into a total of five perspectives. Basically, the defined aspects fully capture the chosen criteria, and therefore, with one exception, it always got a better score compared to other alternatives. The mentioned exception is the visible fluctuation recorded in the criterion "Information ability and equipment", which was given the highest importance by the participants forming the expert opinion of the individual scales (Figure 2, second value).

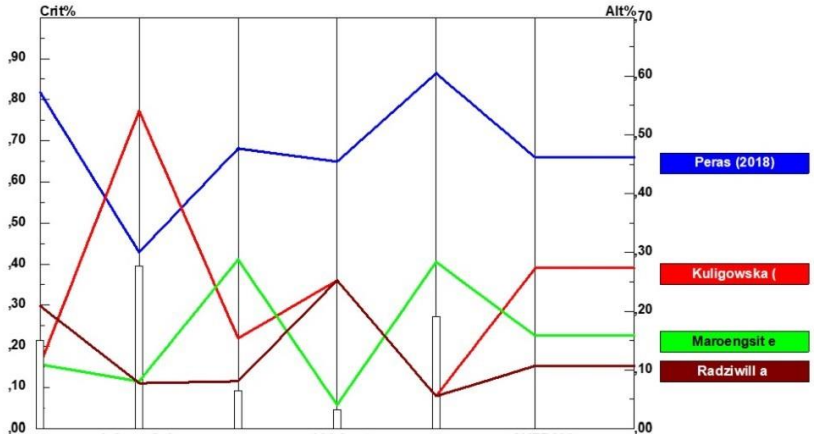

Figure 2. Graphical representation of the selection of a suitable approach to measuring chatbot effectiveness (according to individual criteria - User friendliness, Information ability and equipment, Language level and equipment, Humanity, Business aspect). Source: own processing in the Expert Choice 2000 programme.

In this case, it was surpassed by Kuligowska (2015), who directly deals with evaluating the knowledge base equipment, and in terms of overall results, it finished in second place. In other respects, its measurement method was comparable to the remaining two approaches, and the largest decline was registered in the commercial aspect due to the absence of its consideration. Third place went to Maroengsit et al. (2019), which took one third of what the preferred approach from Peras (2018) and for two criteria was described as unsatisfactory. The approach from Radziwill and Benton (2017) has become the least useful comparison of competitions, which is insufficient from the viewpoint of a total of three criteria and its biggest problem lies in the absence of the implementation method for its proposed attributes and more detailed elaboration into partial metrics.

The 2018 approach from the Croatian author D. Peras was chosen to analyse the approach to measuring chatbot effectiveness.

Acknowledgments: The paper was written with the support of the specific project 2021 grant "DETERMINANTS OF COGNITIVE PROCESSES IMPACTING THE WORK PERFORMANCE" granted by the University of Hradec Králové, Czech Republic and thanks to help of student Andrea Kotoučková.

References

Accenture. (2016). Chatbots in Customer Service. Accenture Interactive: *Part of Accenture Digital*. https://www.accenture.com/t00010101T000000__w__/br-pt/_acnmedia/PDF-45/Accenture-Chatbots-Customer-Service.pdf

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In I. Kompatsiaris, & J. Cave (Eds.), *International conference on internet science* (pp. 377-392). Cham: Springer International Publishing, Lecture Notes in Computer Science, 10673. https://doi.org/10.1007/978-3-319-70284-1

Bureš, V., Jašíková, V., Otčenášková, T., Kolerová, K., Zubr, V., & Marešová, P. (2012). A comprehensive view on evaluation of cluster initiatives. In *8th European Conference on Management Leadership and Governance ECMLG-12* (pp. 74-79).

Business Insider Intelligence. (2016). 80% of businesses want chatbots by 2020. *Businessinsider.com*. https://www.businessinsider.com/80-of-businesses-want-chatbots-by-2020-2016-12

Čermák, P. (2018). *Umělá inteligence: Studijní opora pro kombinované studium*. Olomouc: Moravská vysoká škola Olomouc, o. p. s. https://mvso.cz/wp-content/uploads/2018/02/Um%c4%9bl%c3%a1-inteligence-studijn%c3%ad-text.pdf

Expert Choice. (2020). *The Analytic Hierarchy Process: Structured Decisions: Decision Making for Better Decisions*. https://www.expertchoice.com/ahp-software

FEEDYOU. (2019). *Portfolio*. https://feedyou.ai/success-stories/

Khorozov, A. (2017). Trends Driving the Chatbot Growth. *Chatbots Magazine*. https://chatbotsmagazine.com/trends-driving-the-chatbot-growth-77b78145bac

Kotoučková, A. (2020). Měření efektivnosti chatbota. *Diplomová práce*. Univerzita Hradec Králové.

Kuligowska, K. (2015). Commercial chatbot: performance evaluation, usability metrics and quality standards of embodied conversational agents. *Professionals Center for Business Research*, *2*. https://doi.org/10.18483/PCBR.22

Lokman, A. S., & Zain, J. M. (2010). Chatbot Enhanced algorithms: a case study on implementation in Bahasa Malaysia human language. In *International Conference on Networked Digital Technologies* (pp. 31-44). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-14292-5_5

Maroengsit, W., Piyakulpinyo, T., Phonyiam, K., Pongnumkul, S., Chaovalit, P., & Theeramunkong, T. (2019). A Survey on Evaluation Methods for Chatbots. In *Proceedings of 2019 7th International Conference on Information and Education Technology* (ICIET 2019) (pp. 111–119). https://doi.org/10.1145/3323771.3323824

Minsky, M. L. (1967). Computation: Finite and Infinite Machines. *Prentice Hall Series in Automatic Computation*. Englewood Cliffs.

Molnár, Z. (2000). Efektivnost informačních systémů. *Systémová integrace*. Grada.

Nicastro, D. (2018). 8 Ways to Measure Chatbot Program Success. *CMS WiRE*, Channel: Customer Experience. https://www.cmswire.com/customer-experience/8-ways-to-measure-chatbot-program-success/

Peras, D. (2018). Chatbot Evaluation Metrics: Review Paper. In *36th International Scientific Conference on Economic and Social Development – "Building Resilient Society"*. Varazdin Development and Entrepreneurship Agency (VADEA) (pp. 89–97).

Radziwill, N., & Benton, M. (2017). *Evaluating Quality of Chatbots and Intelligent Conversational Agents*. https://arxiv.org/abs/1704.04579

Rieke, T. D. (2018). The relationship between motives for using a Chatbot and satisfaction with Chatbot characteristics in the Portuguese Millennial population: an exploratory study (*Dissertation)*. University of Porto. Faculty of Economics. https://repositorio-aberto.up.pt/bitstream/10216/116509/2/296743.pdf

Russel, S., & Norvig, P. (2010). *Artificial Intelligence: A modern Approach* (3rd ed.). Prentice Hall.

Shah, N. (2017). Which Is Best for You: Rule-Based Bots or AI Bots? *Chatbots Magazine.* https://chatbotsmagazine.com/which-is-best-for-you-rule-based-bots-or-ai-bots-298b9106c81d

Shawar, B. A., & Atwell, E. (2007). Chatbots: Are they Really Useful? *LDV Forum, 22,* 29-49.

Shieber, S. M. (1994). Lessons from a Restricted Turing Test. *Communications of the ACM*, *37*(6), 70-78. https://doi.org/10.1145/175208.175217

Wakefield, J. (2019). *The hobbyists competing to make AI human*. BBC. https://www.bbc.com/news/technology-49578503