The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots

(Article begins on next page)

06 May 2024

The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots

Authors: Amon Rapp[1], Lorenzo Curti[1], Arianna Boldi[2]

Affiliation: 1 University of Torino, Computer Science Department, Torino, Italy

2 University of Torino, Psychology Department, Torino, Italy

Abstract

Over the last ten years there has been a growing interest around text-based chatbots, software applications interacting with humans using natural written language. However, despite the enthusiastic market predictions, 'conversing' with this kind of agents seems to raise issues that go beyond their current technological limitations, directly involving the human side of interaction. By adopting a Human-Computer Interaction (HCI) lens, in this article we present a systematic literature review of 83 papers that focus on how users interact with text-based chatbots. We map the relevant themes that are recurrent in the last ten years of research, describing how people experience the chatbot in terms of satisfaction, engagement, and trust, whether and why they accept and use this technology, how they are emotionally involved, what kinds of downsides can be observed in human-chatbot conversations, and how the chatbot is perceived in terms of its humanness. On the basis of these findings, we highlight open issues in current research and propose a number of research opportunities that could be tackled in future years.

Contact Author: Amon Rapp

University of Torino - Computer Science Department - C.so Svizzera, 185 – 10149 Torino, Italy

Ph. +393462142386 Mail: amon.rapp@unito.it

1 INTRODUCTION

Recently, we have witnessed a growing interest around conversational agents, software applications interacting with humans using natural language. In particular, over the last ten years text-based chatbots (or simply chatbots), which enable interaction of humans with machines through natural *written* language, have spread in a variety of application domains, so that it has been talked about a "chatbot tsunami" (Grudin & Jacques, 2019). This spreading is driven by flexible platforms that support their design, as well as the possibility of their seamless integration into existing websites and applications (Klopfenstein et al., 2017). Cloud-based cognitive services of composable Artificial Intelligence (AI) building blocks, like the IBM Watson Developer Cloud, can be used by developers to easily build new chatbots, by training the AI to respond to questions posed in natural language about particular intent (Yan et al., 2016).

Given this technological availability, it comes as no surprise that investments in chatbots took off. In 2018, more than 300,000 chatbots were said to be active on Facebook Messenger alone (Boiteux, 2018). Gartner (2017) predicted that by 2022, 85% of customer service interactions will be powered by chatbots. These trends did not go unnoticed by scholars, who started researching how people interact with this technology. Such research has been considered important also because, despite the initial hype and the enthusiastic market predictions, conversing with artificial agents raises issues that go beyond their current technological limitations, rather involving the human side of interaction. Chatbots may not meet the users' high expectations in terms e.g., of language skills, they may elicit a variety of (negative) emotions due to e.g., the frustration of not being understood, and may encourage the formation of biases in the perceptions that users have of the technology (e.g., Forrester Research, 2016; Grudin & Jacques, 2019).

Notwithstanding the proliferation of literature reviews in the field of conversational agents exploiting embodied or spoken forms of verbal interaction (e.g., ter Stal et al., 2020; Clark et al., 2019a), we still know little about research addressed to chatbots exclusively communicating through natural written language. More precisely, we do not have a comprehensive overview of what has been investigated so far from the human side of human-chatbot interaction, i.e., what people expect, feel and, more in

general, experience when they 'encounter' a chatbot. This may be important as chatbot technology cannot exploit 'enhanced' communication means that may enrich the conversation with the agent, like tone of voice in speech-based interfaces, or gestures and face expressions in Embodied Conversational Agents (ECAs). Nevertheless, many studies on person-to-person Computer Mediated Communication (CMC) through text messages (e.g., instant messaging) have consolidated the idea that this form of interaction has unique traits (Werry, 1996) and, despite the lack of cues from body language and vocal tones (Hentschel, 1999), is able to e.g., communicate emotions as in person communication (Derks et al, 2008). What happens, therefore, when people communicate using written messages with another entity, which is not human, is worth to be explored.

In this article, we present a systematic literature review that maps out the current state of research on how humans interact with text-based chatbots. We focus on the last ten years of literature (2010-2020) because during this period both the technology has become mature enough to be accepted by users and proliferation of chatbot research has made the study of user interaction with these technologies increasingly important. In so doing, we adopt a Human-Computer Interaction (HCI) lens, paying attention to the human side of interaction, rather than to performance or capabilities of the system.

Our contributions to HCI are twofold. Firstly, we point out the relevant topics that have been addressed by text-based chatbot research in the last ten years, reporting five main themes that could be relevant for the HCI community. We emphasize that researchers studied whether and why people accept chatbot technology, as well as how trust, engagement and satisfaction may impact on the user experience. We also highlight that the reviewed articles explored how emotions and conversational issues may affect the interaction experience. Moreover, we pinpoint that research paid particular attention to how users ascribe humanness to chatbots.

Secondly, we highlight several open issues in current research and propose a number of research challenges that could be tackled in future years by the HCI community. In particular, we stress that researchers should develop theories capable of accounting for the peculiarities of human-chatbot interaction, carefully define the constructs they use, and better build on previous work. Furthermore, future research should widen the notion of context when exploring the implications of chatbot

technology, conduct studies focused on design, and tackle relevant ethical issues that arise from the users' tendency to ascribe humanness to chatbots.

2 BACKGROUND

2.1 The modern increase in chatbot usage

The chatbot tsunami may be retraced to 2014, when investment in chatbots that could provide more depth on specific topics took off (Grudin & Jacques, 2019). However, it is in the early 2010s that the general public started to get acquainted with conversational agent technology. By 2010, Twitter was flooded with social bots (Grudin & Jacques, 2019), which broadcast, follow, or forward messages generated by other bots or real users (Varol et al., 2017). Although speech-based, Siri brought interaction based on language into the mainstream consumer market in 2010. In 2011, Cleverbot, a chatbot created by Rollo Carpenter, became the first AI program to pass the 'Modified Turing Test' (Gilbert & Forney, 2015), while IBM's Watson was on the game show Jeopardy. While in 2013, WeChat launched its chatbot platform. In sum, in these years, the idea of creating a machine capable of interacting with humans through language, which dates back to Turing's imitation game (Turing, 1950) and the earlier efforts in conversational software, such as ELIZA (Weizenbaum, 1966), PARRY (Colby et al., 1971), and ALICE (Wallace, 2009), became far more popular among the general population.

Albeit all 'chat-bots' may be categorized as interactive bot software, most bots are not conversational: rather, they are automata used for performing tasks such as web crawling and notifying people without meaningfully conversing with users (Grudin & Jacques, 2019). On the other hand, ECAs embody the chat-bot in an animated character and possibly simulate emotions with facial expressions and gestures, thus offering an enhanced form of communication (e.g., Cassel et al., 1999). Likewise, spoken dialog systems move from the textual to the spoken channel, attempting to recognize user speech and offer satisfying spoken answers. In this line, smart or intelligent personal assistants have drawn the attention of the general population in recent years, as big technology players designed

intelligent agents, such as Apple's Siri, Microsoft's Cortana, Google Assistant, and Amazon Alexa, primarily communicating through voice (Clark et al., 2019b).

This is to say that the landscape of bots and conversational agents is complex and its terminology may present terms overlapping and not univocal categorizations. Grudin and Jacques (2019), for instance, proposed a taxonomy of conversational bots identifying three different categories of software on the basis of conversation focus: *virtual companions* engage on any topic keeping a conversation going; *intelligent assistants* take on any topic as well but are aimed at keeping conversations short; *task-focused chatbots* have a narrower range and go deeper, yet brief conversations are their goal. Hussain et al. (2019) identified two main categories of chatbots, task-oriented and non-task-oriented. Følstad et al. (2019) proposed a two-dimensional typology based on duration of relation (short-term vs. long-term) and locus of control (user-driven vs. chatbot-driven) pointing out four chatbot types. Klopfestein et al. (2017) further identified a category of conversational agents called *botplications*, agents endowed with a conversational interface accessible through a messaging platform.

2.2 Research aims

Despite the enthusiastic market predictions, a skepticism toward chatbot technology is growing. In 2016, Forrester Research (2016) emphasized that most chatbots were not ready to handle the complexities of human conversation. It has also been reported that 70% of Facebook Messenger chatbots are not able to answer simple questions (Sullivan, 2017). However, text-based chatbots have already taken hold in many digital services that people now use in their daily life, in domains as diverse as customer service (Følstad & Skjuve, 2019), health (Chaix et al., 2019), entertainment (Jin et al., 2019), and mental wellbeing (Ly et al., 2017). It becomes thus paramount to have a clear overview of how chatbots are really experienced by humans, given also the fact that this technology is relatively new for the general population and their expectations may not be completely met.

Moreover, since text-based chatbots cannot exploit enhanced forms of communication, like speech and gestures, which may make the agents' current abilities/identity more explicit, interaction issues

may become even more pressing. In fact, users of speech-based agents may be easily capable to recognize them as non-human (e.g., Lunsford & Oviatt, 2006), while only-text interaction could be more ambiguous (e.g., Shi et al., 2020). On the other side, CMC research suggests that text-based mediated communication may have unique features and compensate for the lack of visual cues and vocal tones (Hentschel, 1998; Werry, 1996), allowing interlocutors to share emotions and intimacy (Derks et al, 2008). The complexity of this kind of communication, therefore, makes human-chatbot interaction an extremely interesting area of investigation for HCI, entailing the need of outlining the current state of research.

In recent years, scholars have attempted to give order to the different research perspectives on language-based technology by offering a variety of systematic literature reviews on speech interfaces (Clark et al., 2019a), speech-based intelligent personal assistants (De Barcelos Silva et al., 2020), and ECAs (Ter Stal et al., 2020). Other reviews tackled generic conversational agents in specific application domains, such as learning (Abdelkefi & Kallel, 2016; Smutny & Schreiberova, 2020), mental health (Vaidyam et al., 2019), behavior change (Kennedy et al., 2012), and healthcare (Laranjo et al., 2018).

In this systematic review, instead, we focus on a specific kind of conversational agents, namely 'text-based chatbots', *chatbots that are designed to converse and interact with users through natural written language*. Even though such bots have recently become a topic of intense activity, a systematic review of research papers focused on how users interact with them is not available yet.

3 METHOD

To identify relevant themes related to human-chatbot interaction in current research we used the Grounded Theory Literature Review method (Wolfswinkel et al., 2013). This method is rooted in Grounded Theory (Glaser & Strauss, 2017), an approach developed in social sciences that aims to build theories and identify themes across data through an inductive process of data gathering and analysis. What distinguishes Grounded Theory from other approaches is an inductive, rather than a

hypothetical-deductive stance. Grounded Theory has been used either as a way for developing theoretical models emerging from the data, or as a strategy to make sense of a large amount of data, namely a coding method (Gasson, 2009; Mattarelli et al., 2013). In this article, Grounded Theory is taken mainly as a method for analyzing data, in order to add rigor to the process of searching, selecting, and analyzing studies in a review.

In fact, Grounded Theory Literature Review method uses the content of the reviewed articles as empirical material to be analyzed for developing themes through review and has been used by many systematic reviews in HCI (e.g., Nunes et al., 2015; Mencarini et al., 2019). This method follows four different stages: i) Define: selecting the inclusion/exclusion criteria, identifying the appropriate data sources, and defining the specific query to be searched; ii) Search: collecting the papers searched through all the identified sources; iii) Select: defining the final sample by checking the papers against the identified inclusion/exclusion criteria; iv) Analyze: analyzing the selected papers through open, axial, and selective coding techniques. A further stage relates to the presentation and discussion of the papers analyzed.

3.1 Selection criteria

3.1.1 Inclusion criteria

We chose as the unit of analysis for this review 'a study' as defined by Littell et al. (2008, p. 67), namely an "investigation that produces one or more reports on a sample that does not overlap with other samples". In this way, we excluded biases produced by research resulting in multiple reports. We then defined the following inclusion criteria.

i) The articles have to study interaction with chatbots that are exclusively text-based. We included chatbots having a static image for identification, provided that this was not animated nor able to convey dynamic messages through visual language during the interaction (e.g., animated facial expressions).

ii) The articles have to present at least one study with users, in which the technology is studied with reference to the people who interacted with it. This criterion was defined to ensure that the articles did not exclusively focus on system capabilities or performance.

iii) The articles have to focus on the *interaction* between the user and the chatbot with a particular attention to the human-side of such interaction. On the one hand, this entails that the papers should tackle the peculiarities of this interaction, which is text-based. On the other hand, they should present findings on how people experience such interaction, e.g., how users behave, feel, think and react to during the interaction, or how they perceive, expect from, or evaluate the chatbot and the interaction with it. We decided to not circumscribe the selection to papers formally belonging to the HCI field (i.e., published in recognized HCI venues), but to potentially include any relevant research. However, requiring that the papers focus on interaction and how it is experienced by users ensures that they can give insights useful for the HCI community. This criterion guarantees the multidisciplinarity of the reviewed corpus, offering a multifaceted picture of current research that may be relevant for HCI.

iv) The articles have to be published in international journals or in international conference main proceedings and be written in English.

v) The articles have to be published between January 2010 and April 2020. We wanted to analyze research conducted when the technology was not completely unfamiliar to users. Even though, as we said, the beginning of the chatbot tsunami may be retraced to 2014 (Grudin & Jacques, 2019), we decided to also include research conducted in the early 2010s, when conversational agent technology started becoming acquainted by the general public.

3.1.2 Exclusion criteria

As a consequence of the aforementioned inclusion criteria, we excluded those studies that were interested only in assessing the effectiveness of the chatbot in performing specific tasks, without exploring the interaction experience of the user with the chatbot as a conversational technology. For instance, we excluded from our corpus Conversational Recommender System studies (e.g., Mahmood

& Ricci, 2009) that were exclusively concerned about the successfulness of the delivery of the right recommendation to the user (recommendation accuracy). Likewise, we did not include research on chatbots for behavior change (e.g., Pereira & Diaz, 2019) that was exclusively focused on the effectiveness of the behavior change intervention (i.e., whether the behavior changed or not), without exploring interaction matters.

Moreover, we excluded all those papers that conducted user studies only to assess the effectiveness of a particular NLP technology or algorithm (e.g., whether an algorithm for classifying users' intents performed well or not). We also did not include those articles dealing with ECAs, speech technology, or bots that were not able to sustain meaningful conversations. Papers published in the adjunct proceedings of conferences (such as posters, workshop papers, late-breaking result papers) or in books were also excluded.

3.2 Search strategy

The papers presented in this systematic review were collected in April 2020. We chose Scopus, EBSCO APA PsycInfo, and the Association for Computing Machinery Digital Library (ACM DL) as sources for searching relevant articles in human-chatbot interaction. We selected these repositories in order to potentially capture any article addressing relevant topics in human-chatbot interaction (the ACM DL collects computer science papers, PsycInfo psychological papers, and SCOPUS is a more generalist repository). The search was conducted by using the terms and connectors reported in Table 1.

The first part of the query relates to the notion of chatbot including synonyms that could point to text-based conversational agents. The second part refers, instead, to the human aspects of interaction: they combine general and more specific terms that may capture a wide range of topics that may be relevant for exploring the human side of interaction with chatbots. The list of search terms was composed after several iterations and refinement by all the three authors, in line with other HCI research reviews (e.g., ter Stal et al., 2020).

Table 1. Review search terms. Quotation marks are meant to wrap multiple words into one phrase, while asterisks denote truncation to account for alternative spellings.

chatbot* OR "chat bot*" OR "virtual assistant*" OR chatterbot* OR "conversational agent*" OR "natural language interface*" OR talkbot* OR "talk bot*") AND (experience* OR user* OR expectation* OR usability OR understanding* OR misunderstanding* OR bias* OR emotion* OR attitude* OR psycholog* OR interact* OR conversation* OR pragmatics OR cooperat* OR cognit* OR evaluation OR assessment OR social*

In Scopus we limited the search to title, abstract and keywords, to journals and conference proceedings as sources and to articles and conference papers as documents. In PsycInfo we excluded books and dissertations. The search was conducted on papers published from January 2010 and April 2020. These queries retrieved 5194 entries: 3142 in Scopus, 224 in PsycInfo and 1828 in the ACM Digital Library. We exported the results in a table, rearranged the columns with the publications details in order to make them uniform, deleted the duplicates, and removed the papers that clearly did meet the formal criteria (e.g., workshop papers). Eventually, our initial tentative corpus had 4030 papers.

3.3 Article selection

The titles and abstracts of the initial tentative corpus were screened against the eligibility criteria for inclusion in the review by the second author. In this phase, 3741 papers were excluded. The full texts of papers that were considered to be eligible (289) were further assessed to determine inclusion. The full texts were independently screened by the first and the second authors and then discussed. If ambiguities still persisted about the eligibility of a specific paper, input was sought from the third author. At this stage, we identified 78 papers to be included in the corpus. Finally, snowballing was used to screen the references in the included articles using the same technique as used for the screening of the database searches. The final corpus collects 83 papers. Figure 1 shows the flow diagram of the database searches and article screenings.
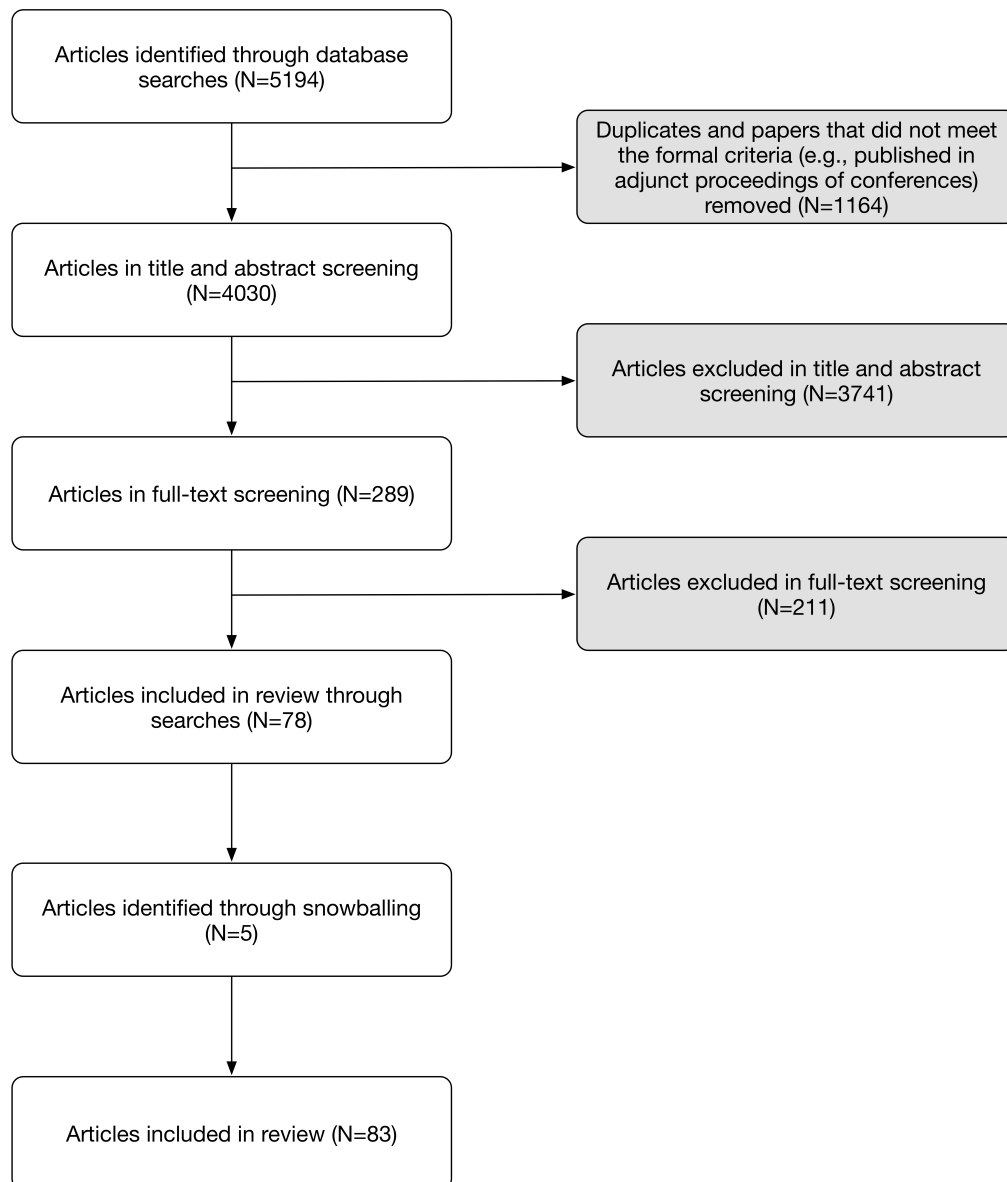
```
┌─────────────────────────────────┐
│ Articles identified through      │
│ database searches (N=5194)       │
└─────────────────────────────────┘
            │                        ┌─────────────────────────────────┐
            │─────────────────────→  │ Duplicates and papers that did   │
            │                        │ not meet the formal criteria     │
            │                        │ (e.g., published in adjunct      │
            │                        │ proceedings of conferences)      │
            │                        │ removed (N=1164)                 │
            ↓                        └─────────────────────────────────┘
┌─────────────────────────────────┐
│ Articles in title and abstract   │
│ screening (N=4030)               │
└─────────────────────────────────┘
            │                        ┌─────────────────────────────────┐
            │─────────────────────→  │ Articles excluded in title and   │
            │                        │ abstract screening (N=3741)      │
            ↓                        └─────────────────────────────────┘
┌─────────────────────────────────┐
│ Articles in full-text screening  │
│ (N=289)                          │
└─────────────────────────────────┘
            │                        ┌─────────────────────────────────┐
            │─────────────────────→  │ Articles excluded in full-text   │
            │                        │ screening (N=211)                │
            ↓                        └─────────────────────────────────┘
┌─────────────────────────────────┐
│ Articles included in review      │
│ through searches (N=78)          │
└─────────────────────────────────┘
            │
            ↓
┌─────────────────────────────────┐
│ Articles identified through      │
│ snowballing (N=5)                │
└─────────────────────────────────┘
            │
            ↓
┌─────────────────────────────────┐
│ Articles included in review      │
│ (N=83)                           │
└─────────────────────────────────┘
```

Figure 1. Flow diagram of the database searches and article screenings.

3.4 Data analysis

Data analysis was conducted on the basis of the key principles of Grounded Theory. All the authors read the selected articles to identify recurrent themes. Initially, each author separately assigned one or more conceptual labels to each article through open coding. We then discussed the coded generated

and defined a final set of open codes. As a second step, these concepts were grouped into conceptual categories through axial coding by each author separately. Finally, the authors discussed together the categories resolving discrepancies and connecting them in a coherent explanatory scheme through selective coding. This step entailed the identification of five main themes, namely, user acceptance, user experience, conversational issues, emotional experience and expression, and humanness, as well as several sub-themes, that will be extensively recounted in the next section.

4 RESULTS

In the following sections we report on the results emerging from the analysis of the 83 papers included in our corpus. The first section provides an overview of where the reviewed articles have been published, while the second and third sections describe the methodological approaches used, highlighting also the systems used, their domains of application, and the typology of users involved. The fourth section focuses on whether and why people tend to accept chatbot technology and also outlines their motivations to use it. Then, we pay attention to how users experience the chatbot, by analyzing three relevant sub-themes related to user satisfaction of, engagement in and trust toward the technology. The sixth section describes conversational issues with the chatbot, namely those interactions that may be problematic or unsatisfying for the user. The last two sections describe research on how people feel and express emotions during interaction, as well as on how they may project humanity into the technology they are using.

4.1 Study characteristics

4.1.1 Publication venues

Papers included in the corpus were published mostly in conferences, which accounted for 57.8% (N=48) of the publication venues. CHI is the most common venue for the papers reviewed here, accounting for 18.1% (N=15) of all papers. Computers in Human Behavior is the preferred journal for human-chatbot interaction research, accounting for 9.6% of articles reviewed (N=8). Publication

venues are scattered across different disciplines, like HCI, psychology, medicine, computational linguistics, and business research. A total of 53% (N=44) of papers were published in HCI-oriented venues. Table 2 provides an overview of where the 83 articles were published.

Table 2. Number of papers reviewed by publication source and year of publication. Conferences can be recognized from the presence of an acronym in brackets.

Venue	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
Conference on Human Factors in Computing Systems (CHI)	-	-	-	-	-	-	-	3	1	4	7	15
Computers in Human Behavior	-	-	-	-	-	1	1	1	1	4	-	8
Human-Agent Interaction (HAI)	-	-	-	-	-	-	-	1	1	1	-	3
International Conference on Internet Science (INSCI)	-	-	-	-	-	-	-	1	2	-	-	3
Artificial Intelligence and Cloud Computing (AICCC)	-	-	-	-	-	-	-	-	-	2	-	2
Behaviour & Information Technology	-	-	-	1	-	-	-	-	-	-	1	2
Communication Studies	-	-	-	-	-	-	-	-	-	1	1	2
Cyberpsychology, Behavior, and Social Networking	-	-	-	-	-	-	-	1	1	-	-	2
International Conference on Human-Computer Interaction (HCII)	-	-	-	1	-	-	-	-	-	1	-	2
ACM Transactions on Interactive Intelligent Systems	-	-	-	-	-	-	-	-	-	1	-	1
Affective Computing & Intelligent Interaction (ACII)	-	-	-	-	-	-	-	-	-	1	-	1
American Society for Engineering Education Annual Conference & Exposition (ASEE)	-	-	-	-	-	-	-	-	-	1	-	1
Applied Computing (AC)	-	-	-	-	-	-	-	1	-	-	-	1
Applied Human Factors and Ergonomics (AHFE)	-	-	-	-	-	-	-	1	-	-	-	1
Cognitive and Computational Aspects of Situation Management (CogSIMA)	-	-	-	-	-	-	-	1	-	-	-	1
Communication Research	-	-	-	-	-	-	1	-	-	-	-	1
Conference on Business Informatics (CBI)	-	-	-	-	-	-	-	-	-	1	-	1
Conference on Human Information Interaction and Retrieval (CHIIR)	-	-	-	-	-	-	-	-	1	-	-	1
Conference on Information and Knowledge Management (CIKM)	-	-	-	-	-	-	-	-	-	1	-	1
Conversational User Interfaces (CUI)	-	-	-	-	-	-	-	-	-	1	-	1
Designing Interactive Systems Conference (DIS)	-	-	-	-	-	-	-	-	1	-	-	1
Digital Health	-	-	-	-	-	-	-	-	-	1	-	1
Education Sciences	-	-	-	-	-	-	-	-	-	1	-	1
European Conference on Cognitive Ergonomics (ECCE)	-	-	-	-	-	-	-	-	-	1	-	1
Human Technology	-	-	-	-	-	-	-	-	-	1	-	1

												Total
IEEE Transactions on Emerging Topics in Computing	-	-	-	-	-	-	-	-	-	-	1	1
IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans	1	-	-	-	-	-	-	-	-	-	-	1
Intelligent Human Systems Integration (IHSI)	-	-	-	-	-	-	-	-	-	1	-	1
Intelligent Tutoring Systems (ITS)	-	-	-	-	-	-	-	-	1	-	-	1
International Conference on Agents and Artificial Intelligence (ICAART)	-	-	-	-	-	1	-	-	-	-	-	1
International Conference on Computational Collective Intelligence (ICCCI)	-	-	-	-	-	-	-	-	-	1	-	1
International Conference on Emerging eLearning Technologies and Applications (ICETA)	-	-	-	-	-	-	-	-	1	-	-	1
International Conference on Entertainment Computing (ICEC)	-	-	-	1	-	-	-	-	-	-	-	1
International Conference on Human Computer Interaction (Interaccion)	-	-	-	-	-	-	-	1	-	-	-	1
International Conference on Human Systems Engineering and Design (IHSED)	-	-	-	-	-	-	-	-	-	1	-	1
International Conference on Information Technology and Electrical Engineering (ICITEE)	-	-	-	-	-	-	-	-	-	1	-	1
International Journal of Emerging Technologies in Learning	-	-	-	-	-	-	-	-	-	1	-	1
International Journal of Human-Computer Interaction	-	-	-	-	-	-	-	-	-	-	1	1
International Journal of Management & Information Technology	-	-	-	-	-	-	-	1	-	-	-	1
Internet Interventions	-	-	-	-	-	-	1	-	-	-	-	1
IT Professional	-	-	-	-	-	-	-	1	-	-	-	1
Italian Conference on Computational Linguistics (CLiC-it)	-	-	-	-	-	-	-	1	-	-	-	1
Journal of Business Research	-	-	-	-	-	-	-	1	-	-	-	1
Journal of Internet Commerce	-	-	-	-	-	-	-	-	-	1	-	1
Journal of Medical Internet Research	-	-	-	-	-	-	-	-	-	-	1	1
Journal of Medical Internet Research mHealth and uHealth	-	-	-	-	-	-	-	-	-	1	-	1
Journal of Medical Internet Research Cancer	-	-	-	-	-	-	-	-	-	1	-	1
Journal of Medical Internet Research Mental Health	-	-	-	-	-	-	1	-	-	-	-	1
Knowledge-Based Systems	-	-	-	-	-	-	-	-	-	1	-	1
Management Science Letters	-	-	-	-	-	-	-	-	-	-	1	1
Methods of Information in Medicine	-	-	-	-	-	-	-	-	1	-	-	1
Mobile and Ubiquitous Multimedia (MUM)	-	-	-	-	-	-	-	-	-	1	-	1
Pervasive Computing Technologies for Healthcare (PervasiveHealth)	-	-	-	-	-	-	-	-	1	-	-	1
Total	**1**	**0**	**0**	**3**	**0**	**2**	**2**	**13**	**16**	**33**	**13**	**83**

4.1.2 Research methodologies

Overall, the most common research methods used in human-chatbot interaction research were quantitative (49.4% of the papers, N=41), while qualitative studies were reported in 10.8% of the papers (N=9). Papers reporting mixed methods accounted for 39.8% (N=33) of the review. Table 7 in the Appendix provides an overview of the methodologies used in our corpus, as well as a summarization of the studies' goals and main findings. In the following we highlight the main methods that have been used across the articles included in our corpus, which are also summarized in Table 3.

Table 3. Research methodologies

Research method	Number of papers
Experiments	33
Conversation log analysis	24
Surveys	22
Interviews	14
Usability tests	9
Field studies	5
Wizard of Oz	5
Interaction log analysis	4
Focus groups	4
Turing Tests and Turing Test-inspired experiments	4

Most of the studies used experiments with a between-subjects design (21 papers), while the others employed a within-subjects design (8 papers), a mixed design (3 papers), and one paper used a within-subjects design for one task and a between-subjects design for the other two tasks of the experiment. These experiments used a variety of measurements tools, like questionnaires, qualitative feedback freely collected at the end of the experiment, open ended questions included in the questionnaire, observation, or interviews.

Survey were also reported in 22 papers. Most of them provided multiple choice questions or Likert scales (e.g., Sanny et al., 2020), while others included a mix of multiple choice and open-ended

questions (e.g., Zamora, 2017), or even elicited exclusively qualitative feedback (e.g., Ta et al., 2020). Some surveys involved a large number of participants: for instance, Xu et al. (2017) recruited 600 participants while Chaix et al. (2019) involved 958 participants. However, the major part of the studies recruited much smaller samples.

Another method widely reported in our corpus was the log analysis, which have been conducted either on the conversations between the user and the chatbot (conversation log analysis) (24 papers), or on the specific interactions that used performed in the system, like the time spent interacting with it (interaction log analysis) (4 papers). The former can be conducted quantitatively, as in e.g., Xiao et al. (2020), or qualitatively, as in Wang and Nakatsu (2013). Often, log analyses have been used as a measurement tool along with questionnaires and interviews in experiments, like in Tärning and Silvervarg (2019), or as a supplementary method in mixed method research, like in Xu et al. (2017). Notably, Li et al. (2020) conducted a conversation analysis on human-chatbot conversation logs, that is an inductive process for analyzing how users' conversations are organized into sequences of actions and systematic practices. This has been performed qualitatively by manually coding 19,451 real-world exchanges (i.e., not induced by researchers but collected from real interactions of users with a chatbot) involving 1,837 users. Analyses on 'texts' can be also conducted on user reviews of chatbots (Ta et al., 2020).

Four papers used focus groups either as a main research method (Medeiros et al., 2019; Mendez et al., 2019), or paired with other methods (Chaix et al., 2019; Kim et al., 2020). Likewise, two studies employed semi-structured interviews as the main method of inquiry (Følstad & Skjuve, 2019; Følstad et al., 2018), while others included this technique along with other quantitative measurement tools.

Five papers used a Wizard of Oz approach, which is an experimental protocol commonly employed in HCI: users interact with a 'simulated' chatbot that is framed as a machine, whereas, in reality, the text showed in interface is produced by a human. This methodology helps to understand specific reactions of humans to a chatbot (Avula et al, 2018; Lee et al, 2020a; Westerman et al., 2019; Cranshaw et al., 2017; Portela & Granell-Canut, 2017). Similarly, four studies were conducted through either a proper Turing Test (Warwick & Shah, 2015) or a Turing Test-inspired experiment (Candello et al, 2017; de

Kleijn et al., 2019; Mori et al., 2019), in which participants attempted to distinguish the nature of hidden entities (human or chatbot).

Nine papers presented usability tests, which is a method used to evaluate the usability of an interface (in this case of a chatbot) and in later years also identified those user studies addressed to evaluate the user experience of a system. This method employs different protocols and measurements, both qualitative and quantitative, like the thinking aloud protocol, in which the user is encouraged to report the difficulties she encounters during the interaction to the experimenter.

As a final note, only a minor quota of papers presented field studies on chatbots, i.e., studies that involved the deployment of a newly developed chatbot in a real context of use without requiring the participants to perform any specific 'artificial' task (e.g., Cranshaw et al., 2017; Kimani et al., 2019; Seering et al., 2020), or that analyzed real-world conversations (i.e., conversations not performed in a laboratory setting) between users and chatbots (e.g., Li et al., 2020; Chaix et al., 2019).

4.1.3 Chatbots' characteristics

The main type of chatbots analyzed in the articles included in our corpus is task-oriented (56.6% of the papers, N=47), which means that these conversational agents are built for helping users to execute a task or solve a problem (as it typically happens in customer service). A total of 26.5% (N=22) of studies focused on conversational or social-oriented chatbots, that is chatbots designed mainly to maintain a good quality of conversations with humans or to establish some forms of relationships with them. In any case, often task-oriented chatbots were designed to keep good conversational flows as well, even though their principal goal was supporting users in accomplishing their tasks. The other chatbots were undefined or presented both the aspects (Table 4).

Table 4. Types of chatbot

Chatbot type	Number of papers
Task-oriented	47
Conversation-oriented	22
With both task- and conversation- oriented aspects	2
Undefined	12

As for the domains of application, the chatbots analyzed in our corpus span from commercial purposes to pedagogical and therapeutic goals. For example, a total of 20.5% of the papers (N=17) focused on chatbots pertaining to the customer service and help-desk domain (e.g., Ashktorab et al., 2019); 14.5% of the papers (N=12) investigated chatbots that are used in healthcare and well-being, three thereof are specifically addressed to mental healthcare (e.g., Fitzpatrick et al., 2017); 7.2% of the articles (N=6) examined pedagogical-educational chatbot (e.g., Tärning & Silvervarg, 2019); 3.6% of the papers (N=3) explored chatbots aimed at providing suggestions (e.g., Jin et al., 2019); 2.4% of the papers (N=2) examined chatbots that are thought to be used in the workplace supporting productivity (e.g., Kimani et al., 2019), while one paper (1.2% of the papers) focused on a Human Resource (HR) chatbot (Liao et al., 2018); two papers dealt with search chatbots (e.g., Avula et al., 2018); one paper focused on a chatbot that is meant to interact during art exhibitions (Candello et al., 2019). This shows how chatbots are versatile and are pervading many aspects of our everyday life (Table 5).

Table 5. Domains of application of the chatbots

Chatbot domain	Number of papers
Customer service/Help desk	17
Health/Well-being	12
E-Commerce	9
Education/Mentoring	6
Small talk	7
Interviews	3
Recommendations	3
Emotional/social support	3
Work/Productivity	2
Search	2

Human Resources	1
Art exhibit	1
Maps orientation	1
Online communities	1
Charity	1
Finance	1
Safety critical environments	1
Group discussion	1
Multiple domains or undefined	11

4.2 Themes

In the following sub-sections, we will describe the themes and sub-themes emerging from the analysis of the reviewed articles, which are also summarized in Table 6.

Table 6. Themes and sub-themes emerging from the analysis.

Themes	Sub-themes	Total number of papers
Accepting the chatbot	-	10
Experiencing the chatbot	Expectation, perception, satisfaction	41
	Trust	
	Engagement	
Conversational Issues	-	6
Emotional experience and expression	Effects of emotions on interaction	16
	Empathy	
	Regulating emotions	
Humanness	Ascribing humanness	25
	Effects of humanness	
	'Pretending' to be human	

4.2.1 Accepting the chatbot

Generally, acceptability refers to a prospective judgment and attitude toward a technology to be introduced in the future, while acceptance points to the judgment and attitude toward a technology after use (Distler et al., 2018; Schuitema et al., 2010). Acceptability/acceptance research often

intertwines with the exploration of the motivations that users may have to use a particular technological artifact (e.g., Pedrotti & Nistor, 2016; Fessl et al., 2011), since in order to accept a technology, or to keep its usage over time, individuals need to be motivated. In our corpus, a total of 12% of the papers (N=10) tackled topics related to the acceptability and acceptance of chatbots. Three papers focus on the assessment of the user acceptance of specific newly developed chatbots, in the teaching (de Medeiros et al., 2019), company (Fiore et al., 2019) and transportation (Zumstein & Hundertmark, 2017) contexts. The other papers, instead, explore wider issues, trying to identify either the reasons why people are open to accept this technology (5 papers), or those motivations that lie behind its use (2 papers).

As for the first line of research, Zarouali et al. (2018) conducted an online survey with 245 participants, who were asked to interact with Cinebot, a chatbot able to assist users in browsing for movies and making movie reservations. Findings reveal that perceived usefulness, perceived helpfulness, pleasure, arousal (the level of excitement in the chat conversation), and dominance (i.e., the extent to which users feel in control) significantly predict consumers' attitude toward the chatbot brand, which determines the likelihood to use and recommend the chatbot. Likewise, Huang et al. (2019) suggest that perceived value plays a major role in affecting behavioral intention to accept chatbots for customer care; whereas Sanny et al. (2020) indicate usefulness, brand image, personality and ease of use as factors that influence the acceptance of chatbots in Indonesia. Nadarzynski et al. (2019), instead, investigated the acceptability of chatbots in healthcare through 29 semi-structured interviews and an online survey with 216 participants (who were not familiar with chatbots). Study results show that curiosity about new health technologies was among the strongest predictors of acceptability. However, a consistent number of participants were skeptical toward these systems, mainly due to concerns about their security and accuracy. Even though they saw chatbots as an anonymous tool to discuss health issues that may be embarrassing, their lack of professional human approach and empathy made them less acceptable. Similarly, Völkel et al. (2020) studied the users' attitudes towards chatbots able to automatically assess users' personality during interaction. Many of the 21 participants involved in an experimental study regarded personality as very sensitive data and

reported that their acceptance towards such systems depended very much on their purpose. Some participants also expressed concerns about protecting their privacy and expected that such a technology could be used for monitoring people.

As for research on motivations to use chatbot technology, Brandtzaeg and Følstad (2017) asked 146 chatbot users from the US to report their reasons for usage. The majority of participants (41%) sought quick and consistent feedback when searching for information or assistance, thus reporting productivity as the strongest motivation. Other reasons were entertainment (20%) and social and relational purposes (10%), i.e., a way to avoid loneliness or fulfill a desire for socialization and curiosity for novel technologies. The authors explain the large preference for productivity as a consequence of the participants' usage of chatbots in the customer service domain, thus being more attentive to efficiency and helpfulness. Another study made up of 24 semi-structured interviews with users of chatbots for customer service substantially confirmed that the promise of efficient and accessible support is the main motivation to use this technology (Følstad & Skjuve, 2019).

To summarize, acceptability and acceptance of chatbot technology appear to be contextual to the specific domain in which the chatbot is used. In healthcare, empathy appears to be relevant to determine acceptability, whereas in the entertainment context, pleasure and arousal may play a major role. Some users, however, seem to be aware of the potential risks that may come from the usage of chatbots, so that potential privacy and security threats could prevent their adoption.

4.2.2 Experiencing the chatbot

Across the various definitions that have been suggested for User Experience (UX) (Desmet & Hekkert, 2007; Hassenzahl & Tractinsky, 2006; Law et al., 2007), there is an agreement in HCI on that it is the result of the interaction between three elements, namely the user, the system and the context (Lallemand et al., 2015). Hassenzahl and Tractinsky (2006, p. 95) define user experience as "a consequence of a user's internal state, the characteristics of the designed system and the context within which the interaction occurs". Even though usability and UX share a wide common ground, the latter may be seen as an expansion of the former, by including additional aspects that intervene when

we interact with a technological artifact: on the one hand, usability proposes an objective approach to interaction, valuing its effectiveness and efficiency; on the other hand, UX is interested in those subjective factors that characterize the experience arising from the interaction with technology (Lallemand et al., 2015). Usability concerns, nonetheless, are commonly included as part of user experience and sometimes are reworked to account for the more holistic approach that UX brings about (Hassenzahl, 2003; Mahlke, 2008; Lallemand et al., 2015)

Concerns about the user experience of chatbots are the most represented topic in our corpus, accounting for 49.4% (N=41) of the papers reviewed. Two papers focus on strict usability matters (Denecke et al., 2020, 2018), while the others tackle wider issues related to the user experience (e.g., Beriault-Poirier et al., 2019; Kimani et al., 2019; Fiore et al., 2019; Cranshaw et al., 2017), also including aspects connected with the overall service experience provided to a 'customer', like branding (Trivedi, 2019) and advertising (Van den Broeck et al., 2019). For instance, Michaud (2018) analyzed 1,023 texts sent to Edward, a chatbot for customer care in hospitality, and concluded that a chatbot must be able to respond appropriately to 'purely dialogue' interactions, such as "Hi" or "Thanks", to maintain an optimal user experience. Candello et al. (2019) analyzed data from an art exhibit where visitors interacted in natural language with three chatbots representing book characters. The authors showed that different audience conditions (such as the presence of strangers) produce some significant effects on the users' self-reported user experience. Holmes et al. (2019), instead, argued that traditional measures for evaluating usability and UX, like System Usability Scale and User Experience Questionnaire, may not be completely accurate when applied to chatbots.

We identify three relevant sub-themes with reference to UX, which pinpoint key subjective aspects intervening in human-chatbot interaction, namely, the satisfaction, the engagement, and the trust that the user may experience. In the following, we extensively describe such sub-themes. Some papers address more than one aspect, so they are reported multiple times.

4.2.2.1 Expectation, perception, satisfaction

A considerable number of papers (16) in our corpus aim to understand and assess the user's 'satisfaction' when interacting with a chatbot (e.g., Liu & Dong, 2019; Mendez et al., 2019; Procter et al, 2018; Galko et al., 2018; Zumstein & Hundertmark, 2017; Jin et al., 2019). HCI regards satisfaction with a technological artifact as a major design goal (e.g., ISO 1998). However, its original definition as "positive attitude towards the product", as part of the standardized usability construct (ISO, 1998), has been lately branded as superficial within a wider UX perspective (Hassenzahl, 2003). In major UX models (Hassenzahl, 2003; Mahlke, 2008), concerns for effectiveness and efficiency were included as instrumental qualities of a system, whereas the notion of satisfaction has been extended to the one of hedonic system's quality (Lallemand et al., 2015). Ortony et al. (1988) define satisfaction as being pleased about the confirmation of the prospects of a desirable event: if a user holds expectations about the outcome of using a certain technology, and these expectations are met, she will feel satisfied (Hassenzahl, 2003). Therefore, when the user's perceptions of the technology (e.g., its perceived value) are equal or larger than her expectation, the technology is satisfactory. When perceptions mismatch expectations, instead, the technology is unsatisfactory and this may negatively impact on the overall user experience (Zhang et al., 2016).

Three of the papers reviewed attempt to identify those factors that may impact the user's satisfaction with specific chatbots, in the luxury (Chung et al., 2018), business (Sensuse et al., 2019), and healthcare domains (Chaix et al., 2019). Two papers, instead, aim to provide more generalizable results, pointing out either those chatbot design features that may increase user satisfaction, or those elements in the conversations that may signal that the user is having a satisfying experience. For instance, an agreeable personality in chatbots for wellness may positively affect satisfaction, more than a conscientious personality (Smestad & Volden, 2018). In Question-and-Answer chatbot interaction, instead, chitchat (asking about chatbot's traits and status, and talking about oneself), frequent occurrences of second-person pronouns (e.g., "how are you"), and casual texting ("do you know" or "tell me") are strong positive signals of playful satisfaction (i.e., satisfaction resulting from playful interaction) (Liao et al., 2018). This agent-oriented interest in conversations, signaled by the

usage of second-person pronouns, is consistent with the tendency to anthropomorphize the agent and engage in chit-chat (Liao et al., 2018), and may indicate that the user has recognized the agent as a 'subject'.

Other papers go in depth in investigating the link between user's expectations and resulting satisfaction. Schumaker and Chen (2010) evaluated six chatbots built to function as general conversationalists, recruiting 347 participants. They found that responses to interrogatives beginning with "Are" and "Where" were the most satisfying. The authors affirm that this is because the answers to these questions are expected to be binary, clichéd, or non sequitur by users, and general and vague chatbot responses fit these types of requests best. In the same vein, by analyzing 215,859 different sessions of interaction with a chatbot of a telecommunication company, Akhtar et al. (2019) noticed that the majority of users expect to receive an answer to the stated question quickly. Otherwise, they leave the chat conversations very soon. The authors also emphasize that knowing users' topics of interests may enable to deliver what they expect at the beginning of conversations increasing their satisfaction.

Følstad and Skjuve (2019), Jain et al. (2018), and Zamora (2017) further explore how users' actual perceptions of chatbots are affected by their expectations, and how such perceptions impact on their satisfaction. The former interviewed users who had recent, real-world experiences with chatbots for customer service. They discovered that participants' perceptions of the chatbots' skills were very much in line with the chatbots' actual capabilities of handling simple requests. It turned out that they did not expect them to have abilities resembling those of a human. This led participants to describe the chatbot as fast, efficient, and helpful, showing that when users hold realistic expectations regarding chatbots' capabilities and their actual perceptions match such expectations, they feel to be satisfied. By contrast, Jain et al. (2018) recruited users with no prior experience with chatbots, asking them to interact with a set of chatbots on the Messenger platform. The log analysis and subsequent semi-structured interviews revealed that participants' expectations were not met by the agents, which turned into a very unsatisfactory experience. Participants were disappointed and frustrated by their mediocre natural language capabilities and also expected the chatbot to retain context across different

sessions, similar to human conversations. These unrealistic expectations made them unable to assess the actual intelligence of the chatbots. As a result, they felt that the chatbots often did not comprehend their intention and were not able to answer them efficiently. Finally, Zamora (2017) asked 54 participants from India and US (only 14 of them had previous experience with chatbots) to interact with different task-oriented chatbots and report on their experience. Many respondents pointed out that their initial expectations were undefined. By interacting with the chatbots, participants then compared the interaction experience to services with which they were familiar, such as search engines, formulating expectations that the chatbot should be high performing, smart, seamless, and personable. Nonetheless, they often felt that information seeking or completing tasks using a standard search engine would be more efficient and informative than the chatbots, which resulted in an unsatisfying experience.

In sum, this research emphasizes that users' expectations affect the actual perceptions of chatbots during interaction, which may influence users' overall satisfaction. Users holding realistic expectations appear to value the actual help that this technology can provide. Those, instead, that expect human-like capabilities tend to focus on the shortcomings of the technology. However, when expectations are undefined, users seem to compare the chatbot with alternative technology services that exist for the particular tasks addressed by the chatbot.

4.2.2.2 Engagement

Albeit there is no univocal definition in literature (Boyle et al., 2012; Doherty and Doherty, 2018), engagement refers to the quality of the user experience that emphasizes the phenomena associated with being captivated by technology (Attfield et al. 2011; O'Brien & Toms, 2008). Engagement has been considered as an affective, behavioral and cognitive association between computer and the user (Goethe et al., 2019), a subjective state where she is completely involved in the activity at hand (Ren, 2016). Be either seen as a core part of the user experience construct (e.g., Saket et al., 2016), or going beyond UX qualities (Lukoff et al., 2018), it appears that engagement points to subjective experiences

in the form of e.g., absorption, involvement, and enjoyment (Liu et al., 2017; Boyle et al., 2012): engaging applications allegedly retain their users over the long term, making them desire to spend time using the application (e.g., Debeauvais, 2016; Brown & Cairns, 2004; Peters et al., 2009; Doherty & Doherty, 2018).

We analyzed eleven papers focusing on the assessment of user engagement (e.g., Shi et al., 2020; Greer et al., 2019; Jin et al., 2019), which highlights the importance of this concept for evaluating the goodness of user-chatbot interaction. This focus also resulted in an exploration of elements that may correlate with engagement. For instance, longer conversations with a chatbot seem to be positively correlated with enjoyment (Cervone et al., 2018), whereas the level of "message interactivity" may predict the level of engagement (Sundar et al., 2016).

Notably, three papers go in depth in exploring those chatbot's features that may enhance the user's engagement, which may span from the use of emojis, to the ability to listen to the user and be timely in providing responses. Fadhil et al. (2018) affirm that a simple question-and-answer task with a health chatbot might require using emojis to engage users when the conversation covers private and personal aspects. In an experimental study with 58 participants, they found that emoji-enriched texts show a lower user's engagement compared to textual messages if the interaction is related to physical well-being, whilst conversations with emojis are preferred by users when the topic is mental being. In a study with 206 participants evaluating an interview chatbot (i.e., a chatbot that aim to draw out users' opinions), Xiao et al. (2020) point out that chatbots with active listening skills, like paraphrasing (restating the user's input to convey understanding), verbalizing emotions (reflecting the user's emotions in words to show empathy), encouraging (offering suggestions to encourage conversation) and summarizing (the key ideas stated by the user) are more effective at engaging users. Avula et al. (2018), instead, conducted a Wizard of Oz study with 27 pairs of participants on the use of a searchbot (i.e., a chatbot that performs specific types of searches) that intervenes dynamically in the conversation during collaborative information-seeking tasks. Results show that users are more likely to engage with searchbots during complex activities and that the time of intervention is the key

for engagement: they engaged with the searchbot because it intervened right as they were about to start searching.

Other two papers prefer to investigate those users' perceptions and characteristics that may affect their level of engagement. In evaluating Vincent, a chatbot for giving and receiving care, with 67 users, Lee et al. (2019) emphasize that the high level of engagement people experienced during the interaction may reflect both socially desirable reactions, such as politeness towards machines as social actors, as well as emotional empathy, i.e., the ability to 'feel for' Vincent. In the same vein, Portela and Granell-Canut (2018) highlight that users' preconception about the chatbot may have an influence on the level of engagement that the user can develop with the chatbot itself. Their study findings pinpoint that users that show an interest in Science Fiction series or films may express concerns about privacy and also dystopic future ideas. Instead, users who work in computer science, or have programming skills, may have skeptic expectations and be less inclined to engage with chatbots. By contrast, people akin to humanities may get surprised about the possibility of engagement with this sort of systems.

It is worth to notice that there may be a trade-off between effectiveness and engagement. An experimental evaluation with 76 users of QuizBot, a chatbot that helps students learn factual knowledge, pointed out that many participants preferred to learn using QuizBot because it was "fun, interactive, and felt like a real study partner." However, this fun conversational side led to inefficiencies in learning: 11.7% of the users' total time in interacting with the chatbot was not spent on learning (Ruan et al., 2019). The authors, nonetheless, affirm that had they removed these casual aspects from QuizBot, users' interest or motivation in learning could be negatively affected.

4.2.2.3 Trust

Trust can be regarded as a cognitive assessment (Mayer et al., 1995), which, nonetheless, also depends on subjective sentiment dependent on the fulfillment of psychological need of security (Frison et al., 2019). Albeit trust and user experience may have separate definitions, trust is often mentioned as a central construct in UX literature (Frison et al., 2019), so that it has been considered a

component of UX (Law et al., 2007; Wright et al., 2003). Väätäjä et al. (2009), for instance, include trust as item of a questionnaire that supports the evaluation of user experience of mobile systems, whereas Rödel et al. (2014) identify trust as an important UX factor in the evaluation of automatic driving systems. Likewise, Frison et al. (2019) confirm that UX and trust influence each other and correlate in the context of automated driving. Traditionally, e-commerce and website research studied the effects of aesthetics and branding on perceived trustworthiness, aiming to enhance it and, consequently, increase the user experience (e.g., Egger et al., 2001; Li & Yeh, 2010; Lindgaard et al., 2011).

Trust seems to be an important factor in human-chatbot interaction as well, especially in those contexts where the conversational agent's actions may have heavy consequences for the user (Zamora, 2017). Six of the papers we reviewed attempt to identify the factors that may affect the user's trust when she interacts with a chatbot. Følstad et al. (2018), for example, asked 13 participants to reflect on experiences of trust toward chatbots for customer experience, using semi-structured interviews. Results show that the chatbots' ability to correctly interpret the users' requests, provide helpful responses, adopt a human-like approach made up of appropriate humor, personality and politeness, and self-present (i.e., to clearly communicate what it can do) are key factors that positively impact on trust. The authors point out that even contextual elements may influence the user's perception of trust, like the brand of the service provider and how it communicates security and privacy policies. This need of security is confirmed by the participants recruited by Zamora (2017), who highlight the necessity to develop trust toward the chatbot when high-risk data and matters of privacy are involved. Yen and Chiang (2020) collected 204 questionnaires and further emphasized that trust in chatbots is predicted by chatbots' characteristics, such as credibility, competence, anthropomorphism, social presence, and informativeness.

Instead, Chattaraman et al. (2019) focused on both chatbots' and users' characteristics, by investigating whether social- versus task-oriented conversation styles of a shopping assistant chatbot influence the trust of older adults with different levels of Internet competency. An experimental study with 121 participants suggests that a chatbot that uses social-oriented interaction style (i.e., it

maintains an informal conversation through e.g., small talk and exclamatory feedback) lead to superior social outcomes (i.e., enhanced perceptions of two-way interactivity and trust) for older users with high Internet competency, while it does not have such effects for older adults with low Internet competency. Likewise, Zhou et al. (2019) try to understand how the personality of a chatbot interviewer and user's personality influence the user's trust in the chatbot. Findings from a study involving 1,280 interviewees reveal that in a high-stakes job interview users are more willing to confide in and listen to (as indicators of trust) a chatbot interviewer with a serious, assertive personality, rather than with a warm, cheerful personality. However, users' personality traits influence perception of trust. For instance, achievement-striving users tend to trust an agent more and such trust makes them eager to impress the agent. Moreover, when a user perceives an agent with a similar personality, she is often inclined to trust the agent more. The authors conclude that designers should create a chatbot that matches users' personality traits, in order to make the conversation enjoyable and effective. The role of users' personal characteristics in affecting trust is pointed out also by Jin et al. (2019), who notice that user with high musical sophistication (i.e., competence about the musical domain) are more likely to feel in control, provide feedback to recommendations, and trust the items recommended by a chatbot designed for providing suggestions about music.

To summarize, the papers included in our corpus show a great interest in assessing commercial and prototype chatbots with reference to UX, addressing aspects like satisfaction, engagement and trust that may enhance the quality of interaction. More precisely, research suggests that satisfaction arises from the interaction between the user's expectations about the chatbot's capability and her actual perceptions of what it can really do. Therefore, it becomes essential to understand how and whether chatbots can meet or modify such expectations. Research also highlights that emojis, active listening skills, timeliness and relevance of the chatbot's responses may be essential elements for enhancing engagement. The goodness of these design features, however, may vary depending on the chatbot's goal and there may be a trade-off between the effectiveness of the chatbot and the engagement it is able to elicit. Finally, reviewed research suggests that chatbots' characteristics, such as an assertive personality, a transparent self-presentation, or a social-oriented interaction style, may be beneficial to

trust in certain contexts and for certain users. In fact, both engagement and trust appear to be affected by the user's perceptions and characteristics. Engagement, for instance, may be dependent on the user's preconceptions about the chatbot, while trust may be influenced by the user's personality and competence.

4.2.3 Conversational issues

Interaction with chatbots can be problematic. Errors and biases both from the human side and from the conversational agent side appear to be quite common in human-chatbot conversations, which may produce misunderstandings or irreversible breakdowns. In this section we analyze those papers that investigate the conversational problems that may arise during the interaction with a chatbot, which accounted for 7.2% (N=6) of the review. These may be split in two groups: three papers examine users' 'bad language', typos and errors, which may prevent the chatbot from understanding the ongoing conversation; other three papers, instead, focus on the development of human-chatbot conversations, paying attention to those issues that may hinder their progress and successful evolution, as well as those user strategies that can potentially fix them.

As for the first line of research, Hill et al. (2015) analyzed 100 Cleverbot conversations and 100 Instant Messaging conversations from students finding that people were significantly more inclined to use profanity words in the chatbot messages compared to human–human CMC. Moreover, it emerged that users used more pronouns, swear words, social words, negative emotion words, and sexual words when communicating with Cleverbot as opposed to another human. Similarly, Park et al. (2018) examined the conversations of 58 college students with Rose and Mitsuku, two popular chatbots available online. They discovered that the readability of the students' texts was on the equivalent of a 5th-grade level rather than college level, and that they made a variety of spelling and grammatical errors. For instance, many comments had a lack of capitalization at the beginning of sentences and there were many missing apostrophes in contractions. These, in certain cases, may entail the impossibility for the chatbot to understand what was meant. In the same vein, Wang and Nakatsu (2013), in analyzing 1,500 users' interactions with Confucius, a conversational agent that mimics

philosopher Confucius, found that 11.2% of the examined users' utterances contained obscene language, random keystrokes, nonsenses, and writing mistakes, while 4.2% was factoids aimed at testing the chatbot skills.

As for the second line of research, Li et al. (2020) studied conversational non-progresses by analyzing 2,597 users' conversations with a banking chatbot through conversation analysis. The analysis shows that conversations do not make progress much more often when the user requests information (68%) than when she provides it (12%). Moreover, non-progresses mainly occur when the chatbot does not correctly recognize the user's intent, or the content inserted by the user is unexpected by the chatbot (e.g., staying in a conversational topic after the chatbot had moved on to a new topic). Users, however, adopt different coping strategies from trying to quit (65%) (e.g., temporarily changing subject, or abandoning the chatbot service) to reformulating the message (35%) (e.g., rephrasing, repeating the same words, asking a new topic on the same subject). Likewise, Ashktorab et al. (2019) investigated users' repair strategy preferences for conversational breakdowns with chatbots by conducting a pairwise comparison experiment with 203 participants. Findings show that participants largely prefer 'options' strategy, in which the chatbot provides options of potential intents in which it has the highest confidence, thus restricting interaction within its capabilities. Other preferred strategies are highlighting words that the bot did not understand or providing explanations of the chatbot understanding in a confirmation message. These strategies were preferred because they explicitly acknowledge a potential breakdown, manifest initiative from the chatbot, and are actionable to recover from the breakdown. Zaroukian et al. (2017), instead, identify an automation bias that may prevent the conversation with a chatbot to be successful. This phenomenon occurs when the chatbot misunderstands the user's request, providing her with an incorrect answer, and, nonetheless, the user exhibits a complacency behavior. On the basis of an experiment with 161 participants, the authors suggest that automation bias occurs because users tend to accept the agent interpretation when it is generally correct.

In sum, the papers recounted in this section show that humans, in written dialogue with chatbots, may use poor language with sentences riddled with spelling and grammar errors, obscene words, and

nonsenses, which could undermine the performance of the chatbot. During the conversation, chatbots may also fail in understanding the user's intent or provide irrelevant responses. In such cases, users prefer that the chatbot explicitly acknowledges its misunderstanding, taking the initiative in providing repairing solutions. However, users are able to adopt different coping strategies when they encounter conversational issues, which may span from temporarily changing subject to reformulating the message.

4.2.4 Emotional experience and expression

Emotions are a fundamental aspect of human-human communication as well as of human-machine interaction: any interface that ignores a user's emotional state or fails to show the appropriate emotion may risk being perceived as cold, socially inept, untrustworthy, and incompetent (Brave & Naas, 2002). Many of the articles included in our corpus highlight the importance of emotions in human-chatbot interaction, accounting for 19.3% (N=16) of the review. Xu et al. (2017), for instance, analyzed a sample of two hundred user's requests to chatbots on social media, finding that more than 40% of them are not informational but emotional, i.e., in which users intend to express an emotional state.

More precisely, this literature review shows three main ways in which emotions have been addressed by chatbot research: by investigating what kinds of emotions are generated and expressed during the ongoing conversation with the chatbot (5 papers); by focusing on empathy, as a fundamental emotion to develop a close connection with the agent (6 papers); by seeking strategies for regulating the emotions of the user (4 papers).

4.2.4.1 Effects of emotions on interaction

Five papers explore what kinds of emotions stem from the conversation and their effects on interaction. By and large, users' negative emotions emerging during interaction seem to be the main culprit for abandoning the conversation, whereas positive emotions not only play a major role in

preventing communication breakdowns but may also improve the outcomes of a technology intervention. Tärning and Silvervarg (2019), for instance, studied fourth grade students' reactions to a pedagogical agent aimed at supporting learning outcomes, pointing out that responses to the digital tutees' feedback, competence, and attitude were emotionally loaded: students with low self-efficacy who showed a positive disposition toward the chatbot, e.g., expressing happiness regarding its abilities, performed better in learning tasks. By contrast, Wang and Nakatsu (2013) suggested that 'bad' emotions, like unhappiness or frustration, may be elicited by irrelevant responses provided by the chatbot, leading to drastic worsening of the conversation or even communication breakdown.

Perez-Marin and Pascual-Nieto (2013), instead, highlighted that the chatbot's 'mood' itself may impact on the willingness of users to continue the interaction. The authors investigated children's responses to four different chatbots with different personalities and moods: even though children did not express any sign of aggressiveness toward the bad-tempered chatbot, they spent least time with it, rather preferring the other agents. The chatbot's ability of expressing emotions also appears to influence users' self-disclosure behavior, namely the gradual unveiling of personal information, thoughts and feelings. Lee et al. (2020b) ran a study with 47 participants interacting with chatbots with different chatting styles for three weeks. They discovered that computer agents' self-disclosure could facilitate the users' self-disclosure. In addition, chatbot's self-disclosure had a positive impact on participants' perceived intimacy and enjoyment with the chatbot. Similarly, Lee et al. (2020a) conducted a laboratory experiment with 64 subjects and found that when the language used by a chatbot expresses emotional states, the chatbot draws users' cognitive attention to the social aspect of their interaction partner, producing increased co-presence.

4.2.4.2 Empathy

A second group of six papers focus on empathy as a 'special emotion' that chatbots may be able to express, and which may consistently change the quality of interaction especially in terms of involvement and relationship development with the chatbot. Empathy is commonly understood as a

two-dimensional construct, whereby cognitive empathy concerns the comprehension of another person's emotions (Vossen et al. 2015) and affective empathy relates to an emotional response to the perceived emotion of others (Mehrabian & Epstein, 1972).

Liu and Sundar (2018), for instance, show that when we ask a chatbot for advice about a health problem its expression of affective empathy is perceived by users as more supportive than simply providing medical information. Fitzpatrick et al. (2017) developed a self-help chatbot for college students with symptoms of anxiety and depression, emphasizing its nonhuman nature by calling it Woebot. They found that users valued the chatbot's empathic responses and suggested that therapeutic relationship can be established even between humans and agents that are purposefully designed as nonhuman, provided that they are capable of expressing empathic emotions. Likewise, Ta et al. (2020) suggest that chatbots may be a promising source of everyday companionship, providing emotional support and increasing positive affect through uplifting and nurturing messages. Based on the analysis of 1,854 user reviews of Replika, a popular companion chatbot, and 66 questionnaire responses, the authors stress the emotional support that the chatbot can provide, including expressions of care, love, and empathy, which may lead users to feel less lonely. Portela and Granell-Canut (2017) compared users' interaction with a chatbot to interaction with a human pretending to be a chatbot, following a Wizard-of-Oz methodology. The human was meant to provide more empathic responses by referring to personal topics. The authors discovered that users were strongly emotionally involved when the chatbot showed empathic behaviors, which provoked on them the same empathic feeling.

Urakami et al. (2019), instead, conducted a survey with 150 novice users of chatbots, finding that statements expressing help or showing interest were evaluated as highest in regard to empathy by users. However, some participants reported that they felt troubled, irritated and annoyed when a machine expresses emotions or pretends to understand their feelings: the authors conclude that the user's belief that chatbots are 'only machines' and are thus expected to behave unemotionally may affect how a system expressing empathy is perceived.

Following a slightly different research line, Lee et al. (2019) compared two versions of Vincent, a self-compassion chatbot that people could care for (care-receiving Vincent) and be cared by

(caregiving Vincent). Compassion builds on empathy, allowing individuals to relate to sufferers in a healthy way, without over-identifying with them, which may lead to vicarious pain. The authors note that caring for a chatbot can enhance one's own self-compassion, improving her well-being and mental health.

4.2.4.3 Regulating emotions

The last group is composed of four papers that investigate how the chatbot can influence the individuals' emotional states over the long term by proposing novel prototypes. SERMO (Denecke et al., 2020) is a chatbot that regulates the user's emotions based on cognitive behavioral therapy, asking daily the user about her mood and offering regulatory activities such as mindfulness exercises. The psychologists and psychotherapists involved in the user study emphasized that the chatbot could be able to encourage patients to elaborate on their emotions, especially those individuals that have difficulties in expressing themselves in a face-to-face encounter. Similarly, Vivibot (Greer et al., 2019) delivers a cognitive and behavioral intervention to increase positive emotion in young adults who have undergone cancer treatment. Participants of the user study perceived the interaction with the chatbot as nonjudgmental, positively affecting their receptivity to learn skills to manage negative emotions. Medeiros et al. (2019) also designed a chatbot that is able to recognize the user's stressful situations and suggests different emotion regulation strategies. The preliminary evaluation with 6 participants highlighted that conversations about stress-related topics should be driven by the user, while the chatbot should not start talking about user's negative emotions on its own initiative. Finally, Kuramoto et al. (2018) present a conversational agent aimed at regulating workers' anger in a customer support working environment. The evaluation with 18 participants pointed out that the chatbot itself, with its mere presence, may produce a high level of anger, even though it was able to regulate the worker's emotion by showing sympathy or enacting strategies of emotion suppression. This points out the double-edged effects that chatbots may have on emotions.

To summarize, the papers included into this theme highlight the complex ways in which emotions are involved in human chatbot-interaction. Different emotions elicited by the chatbot may provoke people in different ways: while positive emotions emerging during the conversation with a chatbot can be an indicator of the successfulness of the interaction, negative emotions need to be addressed as they may entail a communication breakdown. Studies also emphasize that it is important to encourage both chatbot's and user's self-disclosure and empathy. While self-disclosure may increase the perceived intimacy and enjoyment of the interaction, as well as the sense of co-presence, empathy can produce more engaging interaction and favor the establishment of a closer connection between the user and the chatbot. However, users' expectations and beliefs could have an impact, as individuals who expect that the chatbot should behave as a machine may interpret its empathy behaviors as negative. Finally, chatbots seem to be able to affect users' emotions even over the long term, being nonjudgmental and favoring the user's emotional expression and rehash. However, research also highlights that they may produce side-effects, as their mere presence may provoke changes in the user's emotional states going against the intended goal of the technological intervention.

4.2.5 Humanness

Research on humanness has typically focused on robots and embodied conversational agents, pointing out how embodied features, like the agent's appearance, its gestures and facial expressions affect whether and how users perceive humanness in the agent (Ho & MacDorman, 2017; Schwind et al., 2018). Several works also explored conversations with intelligent personal assistants which emulate human aspects of speech (Gilmartin et al., 2017). This review points out that humanness is a central topic in research that studies human-chatbot interaction as well. Some of the studies presented in this section start from theories developed to account for interaction with robots, like the uncanny valley theory, which suggests that perceptual difficulty in discerning a human-like object will evoke negative feelings (Mori, 1970). More precisely, the 25 papers presented in this section, accounting for 30.1% of the review, can be divided into three different lines of research: i) those that study how people ascribe humanness to chatbot, ii) those that explore what kinds of effects such humanness may

produce on the interaction experience, and iii) those that propose chatbots in which humanness is a pivotal design element.

4.2.5.1 Ascribing humanness

A first line of research, made up of 9 papers, attempts to give answers to questions like: what does humanness constitute in a conversation? What are the essential characteristics that chatbots should exhibit to appear human-like agents? Is humanness always welcomed by users? This research explores those linguistic, psychological, and communicative aspects that people consider during a conversation in order to figure out whether they are talking to a human or a machine.

A first group of four papers focus on the 'formal aspects' of interaction, such as grammar and language, as predictors of perceived humanness. De Kleijn et al. (2019) investigate whether there are unique features of human language perceived by humans as humanlike. Fifty-three participants rated the humanness of a set of sentences based on questions and responses of real chatbots, which were manipulated for grammatical construction. The results indicate that center-embedded sentences are perceived as less humanlike than right-branching sentences and more plausible sentences with regard to world knowledge are regarded as more humanlike. Westerman et al. (2019), instead, analyzed the effects of 'typing' on 83 users. Whereas capitalized words did not influence their perceived humanness, typos made the chatbot much less human-like. Although previous literature reviewed by the authors made them hypothesize that typos would humanize a bot by making it seem less perfect (thus, more human), their data suggest the opposite. Araujo (2018) explored the extent to which human-like cues such as language style and giving the chatbot a name can influence perceptions of humanness. An experiment with 175 participants showed that the usage of human-like language or name are sufficient to increase perception of the agent as being human-like. If grammar, typing, language style and name appear to have an impact on users' perception of humanness, typefaces do not seem to have the same effect. Candello et al. (2017) used a blind Turing test-inspired approach with 199 participants to show that machine-like typefaces (like OCR) bias users towards perceiving

the adviser as machines but handwritten-like typefaces (like Bradley) have not the opposite effect. However, previous knowledge and exposure to artificial intelligence chatbots seem to bias users to perceive chats as created by machines.

Two other papers try to understand what 'psychological' factors people consider to ascribe humanness to a conversation. Warwick and Shah (2015) ran an imitation game (i.e., a Turing test) involving human interrogators attempting to ascertain the nature of hidden entities (human or chatbot) during 5-minute unrestricted questioning period. The chatbot's goal was to cause the human interrogator to make a wrong identification, whereas the role of the hidden humans was to be themselves. The authors recruited 30 different interrogator-judges in the test, who interacted with five different chatbots and 25 hidden humans. The test shows that judges base their own judgments on the interlocutor's capabilities of lying and being humorous, as well as on the presence of a common knowledge ground between the hidden entities and themselves. In the same vein, Mori et al. (2019) studied how humans identify human-likeness in online text-based Question & Answer communication, by examining whether 30 users could correctly identify text created by a chatbot and text created by a human in a Turing-like test. Results show that the chatbot-created text left negative impressions such as "businesslike and cold" and "questions and answers that don't match up" and that these impressions were factors in distinguishing chatbot-created text from human-created text.

Finally, three papers pay more attention to the dynamic of interaction, identifying those conversational styles that may make a chatbot appear more human-like. Go and Sundar (2019) conducted an experiment with 141 participants, showing that a high level of conversation contingency (i.e., responses exchanged are interconnected with each other, as it happens in human-to-human conversations) compensates for the impersonal nature of a chatbot that is low on anthropomorphic visual cues. Moreover, identifying the agent as human (i.e., because it is named like a human) raises users' expectations for contingent conversations, as it triggers a cognitive heuristic about the essential characteristics of humanness. The authors also notice that a mismatch between the identity cues and the visual cues may lead to negative outcomes. For instance, participants identifying the agent as a chatbot (because it is named like a chatbot) may have made them expect machine-like cues, but the

use of a highly anthropomorphic visual cue may have been perceived as disingenuous resulting in lower perceptions of homophily and less favorable attitudes towards the chatbot. Svenningsson and Faraon (2019), instead, conducted a survey with 69 participants, finding that making a conversation with chatbots human-like is not always preferred by users, as chatbots should avoid small talk, interact in a clear, concise way, not involve humor, and identify themselves as chatbots. The authors note that the results could be context dependent as features like small talk may not be welcomed in a service-oriented context but may be accepted in a social context. Finally, Morrisey and Kirakowski (2013) carried out an experimental study and a survey with 14 and 20 participants respectively to identify dimensions on which users might judge the naturalness of a chatbot, i.e., whether it is able to conduct human-seeming conversations. These resulted into conscientiousness (e.g., the ability to keep track of the conversation), originality (e.g., taking initiative during the conversation), manner (e.g., developing a relationship with the user), and thoroughness (e.g., using appropriate grammar).

4.2.5.2 Effects of humanness

A second line of research is interested in studying the effects of chatbots' human-like features on users. Specifically, a group of four papers rely on theory of uncanny valley to investigate whether human-like chatbots may engender negative feelings. This theory suggests that cues in non-human agents that encourage attribution of mental abilities that are allegedly deemed to be distinctively possessed by humans may trigger eerie sensations among users, because their mental categorization conflicts and threats to human distinctiveness (Stein & Ohler, 2017). Other nine papers, instead, focus on the "social effects" of human-like chatbots, highlighting how perception of humanness may increase the willingness to establish a common ground between the chatbot and the user and to express more sociable desirable traits, as well as intimacy and closeness.

As for the first group, in Liu and Sundar (2018), it emerges that the expression of empathy and sympathy, an affective process involving feelings of sorrow, concerns, and sadness for other's distressful events, are essential elements for ascribing humanness to a chatbot. However, two

experiments with a total of 246 participants show that, for those individuals who believe that 'robots' can have real feelings, chatbot's expression of sympathy and empathy may be spine-tingling, according to the theory of uncanny valley. This goes in line with the results found by Ta et al. (2020) in their study on Replika, which highlighted that certain users may be unsettled by the chatbot's ability to sound and interact like a real human.

By contrast, in an experiment with 28 participants aimed at investigating whether the lack of transparency regarding a chatbot's nature may provoke uneasy feelings, as the uncanny valley theory suggests, Skjuve et al. (2019) found that the state-of-the-art in a humanlike chatbot (i.e., Mitsuku chatbot) seems not to elicit uncanny feelings, even when there is uncertainty related to its true nature. People may think that there is something strange about the conversation, but they seem not to perceive it as uncanny. Likewise, Ciechanowski et al. (2017) conducted an experiment with 31 participants studying whether either a chatbot with an avatar that is graphically very similar to a human, or a text chatbot without any avatar image, may engender uncanny feelings. The authors did not find any uncanny valley effect in the condition with a text-based chatbot. This effect, however, was identified in the avatar-based condition, leading users to perceive the chatbot more inhuman, weirder, and less competent that the only-text chatbot.

As for the second group of papers, Corti and Gillespie (2016) show how construction of intersubjectivity in human-chatbot interaction may be affected by perceived humanness. An experiment with 108 participants demonstrates that when people speak to an artificial agent under the same conditions as everyday human-human interaction (i.e., a human that is assumed to communicate autonomously), they more persistently try to establish common ground (i.e., they exert more intersubjective effort) relative to conditions wherein knowledge that an interlocutor's words are determined by an agent is explicit and/or the interface is nonhuman (i.e., a text-based chatbot). This finding is important because it emphasizes that the mere knowledge of a chatbot being something artificial might suppress the intersubjective effort to repair of misunderstandings people exert when interacting with it. Similarly, Beattie et al. (2020) conducted an experimental study with 96 participants to explore the role of emojis in engendering perception of humanness. Results show that

users rated emoji-using chatbot message sources similarly to human message sources, and higher than verbal-only messages on social attraction, CMC competence, and credibility. Likewise, Følstad et al. (2018) found that a more human-like conversational style (e.g., a personal style with some humor when appropriate) would be beneficial to building trust between the user and the chatbot. This goes in line with the findings reported by Ly et al. (2017), who evaluated a chatbot for mental well-being and discovered that perceiving the chatbot as somewhat "a (real) person" encouraged the development of a bond between the user and the chatbot.

In the same vein, Mou and Xu's (2017) experiment makes visible that users tend to be more open, more agreeable, more extroverted, more conscientious and self-disclosing when interacting with humans than with a chatbot. The findings suggest that believing that the interlocutor is human may lead users to demonstrate more socially desirable traits in communicating. Pérez-Marín and Pascual-Nieto (2013) further suggest that children tend to perceive the chatbot as a human and a friend. This believed humanness may lead children to exhibit more prosocial behaviors, reducing their aggressiveness toward the chatbot. Lee et al. (2020a) focus instead on intimacy and closeness in a laboratory experiment with 64 participants. They discovered that the more users perceive the chatbot to have a mind, the stronger a sense of co-presence and closeness they report. Shi et al. (2020) conducted an experiment with 790 participants to be persuaded by a chatbot for charity donation, exploring whether the perceived identity of the chatbot (human vs. bot) had significant effects on the persuasion outcome (i.e., donation) and interpersonal perceptions. They observed higher donation probability when the perceived identity was human. Moreover, they found that when participants thought they were talking to a human, they thought that the partner was more competent, confident, sincere, and warm, and the conversation was also deemed more engaging, natural, and persuasive. Interestingly, the authors also discovered that some people still perceived an identified bot as a human.

However, there may be also dark sides in increasing the humanness of chatbots with reference to social behaviors. On the basis of Computers Are Social Actors (CASA) theory, which suggests that people apply human social rules and expectations when interacting with computers (Lee & Nass,

2010), Schuetzler et al. (2019) suggested that design elements that make chatbot interactions more human-like may induce undesired strategic behaviors from human deceivers to mask their deception. This recommends that for applications in which it is desirable to detect when users are lying, like automated screening systems for border crossings and job interviews, the pursuit of human-like interactions may be counter-productive.

4.2.5.3 'Pretending' to be human

The last line of research is represented by three papers and relates to the attempts to design chatbots that encompass human-like features. Kim et al. (2020) designed GroupfeedBot, a chatbot that facilitates group discussions. A qualitative study revealed that users describe the chatbot using personal pronouns ("he" or "she") and as e.g., a "manager," "supporter," or "moderator". Moreover, participants reported that the chatbot served as a group member and "felt really like a human". Seering et al. (2020), instead, created a social chatbot and then studied it in an established online community on Twitch. The chatbot was characterized as a child learning how to talk and behave. It was also designed to interact with the community members in ways that changed both through acquisition of new vocabulary and through 'aging' through pre-designed age stages. The chatbot triggered parenting behaviors, whereby a number of users attempted to teach BabyBot appropriate manners and also created deep relationships with it over time. As time went by, the community as a whole welcomed the chatbot as a member of the community, introducing new members to it in a way that described it more as an agent than an artificial entity. Finally, Calendar.help (Cranshaw et al., 2017) is an agent that provides fast scheduling specifically designed to respond like a human: users interact with the system via email, delegating their scheduling needs to it as if it were a human personal assistant. The tasks that it cannot execute automatically are listed on an in-house online platform similar to Amazon Mechanical Turk, in which human workers handle the task for it: for instance, if it fails to classify whether a message is related to a certain meeting, a human provides this classification. This architecture also entails that the user never knows exactly who – or what – she is

communicating with when she emails the agent. The evaluation in the wild conducted with 178 participants showed that many users thought of Calendar.help as a real person.

To summarize, the papers presented in this Section show that people may consider different dimensions when ascribing humanness to chatbots, as well as rely on different cues to infer whether their interlocutors are humans or machines. Beyond the linguistic level, in which individuals pay attention to grammar, typing, plausibility, and language style, they also value certain psychological and interactional factors, like capabilities of lying, being humorous, and a high level of message interactivity, as well as consciousness, originality, manner, and thoroughness. By and large, these studies point out that embodiment is not a precondition for an agent to be perceived as human-like (Araujo, 2018), as text-based chatbots can be ascribed with humanness on the basis of their (written) linguistic behavior.

This said, considering chatbots as humans may provide both positive and negative outcomes. On the one side, chatbots that appear more human-like may foster users' prosocial behaviors, like willingness to repair of misunderstandings, trust, tolerance, sense of co-presence, closeness, and predisposition to donation. On the other side, an increased human-likeness may encourage users to perform deceptive behaviors and engender feelings of eeriness, as the theory of uncanny valley suggests. As for this point, research presents contrasting results, showing that uncanny valley might be sensitive to individuals' beliefs (Liu & Sundar, 2018).

Nonetheless, research shows that human-like chatbots also entail emotional aspects that are most peculiar to human-human interaction, like empathy and self-disclosure. BabyBot (Seering et al., 2020) is designed to behave like a baby human and this elicits feelings of care for the chatbot. In parallel, care-receiving Vincent is able to increase self-compassion "most likely through a significant change in participants' sense of common humanity" (Lee et al., 2019: 703). Likewise, believing that the chatbot is human (Mou and Xu, 2017), for instance because it is able to self-disclose (Lee et al., 2020b), may lead users to self-disclose too. In other words, these examples show how specific emotions (like empathy and compassion) and emotional processes (like self-disclosure) are strictly connected with our sense of humanness. Self-disclosing chatbots and care-receiving chatbots also

point to an underserved interaction paradigm: here the chatbot is not merely meant as 'serving' humans, i.e., as a means or a 'tool' for achieving a goal like providing support or executing a task; rather, it becomes the 'subject' of an intersubjective relation, in which both the parts of the interaction have the same dignity and their own 'needs' to satisfy.

5 DISCUSSION

This review highlights the characteristics of empirical studies on human-chatbot interaction conducted in the last ten years. Interestingly, most research on users interacting with them is extremely recent (more than two thirds of the reviewed papers are published from 2017). This reflects the spreading of commercial chatbots in the last few years and signals the growing interest for this technology in the academic world. Methodologically, papers tend to study the interaction with chatbots mainly quantitatively by conducting experimental studies driven by specific hypotheses. However, only four studies (e.g., Ly et al., 2017) conducted a power analysis to determine the sample size: this is a limitation of the quantitative work that we reviewed.

The review also signals a lack of studies that exclusively employ qualitative techniques (such as the conversation analysis performed by Li et al., 2020), which could be useful to gain rich insights on the interaction with chatbots. We identified five main themes across the papers we reviewed. Based on these findings, in the following we emphasize several key challenges and research opportunities that could by tackled by the HCI community in the future.

5.1 Developing theories

Our review highlights that the papers in our corpus both explored theory-inspired research questions and produced theoretical knowledge on how people experience the interaction with chatbots. A consistent number of papers embrace theoretical perspectives that have been developed for other technology domains or that were thought as general theories of technology. For instance, the uncanny valley theory, which is employed by several papers in our corpus (e.g., Liu & Sundar, 2018; Ta et al.,

2020; Skjuve et al., 2019; Ciechanowski et al., 2017), was originally addressed to understand the interaction with robots. By contrast, CASA theory, which grounds several reviewed studies (e.g., Schuetzler et al., 2019; Mou & Xou, 2017; Lee et al., 2020a), has been thought to be applied to every digital technology.

These studies are certainly useful to explain relevant aspects of human-chatbot interaction. Nonetheless, we may identify a main limitation in the theoretical knowledge produced by this research. Research often missed to develop theorizations capable of accounting for relevant *peculiarities* of text-based interaction (if any) with artificial agents. This parallels the lack of theories for understanding user language choices in speech interface interactions (Clark et al., 2019a). Peculiarities of text-based interaction could be likely better identified and explained by research designs that compare chatbots with other similar conversational technologies, such as ECAs or speech-based agents, and by referring to theories that address the written nature of the mediated communication with chatbots.

Even though several papers explore the differences between human-human communication and human-chatbot interaction (Mou & Xou, 2017; Hill et al., 2015; Warwick & Shah, 2015; Corti & Gillespie, 2016), there is a shortage of studies that comparatively identify the distinctive traits of chatbot technology, which is exclusively text-based, and theoretically account for them. For instance, this review signals that text-based chatbots may be capable of eliciting a sense of humanness even without providing the user with any 'sensorial stimuli'. It is sufficient that the chatbot is named as a human (Araujo, 2018; Go & Sundar, 2019) or uses emojis (Beattie et al., 2020) to elicit the user's perception of humanness. Moreover, some people still perceive as a human a chatbot that self-presented itself as a machine (Shi et al., 2020). This contrasts research on speech-based conversational agents, which highlights that users define the qualities of speech interfaces differently to humans, clearly perceiving differences in e.g., expressiveness and paralinguistic features, as well as in emotional and social behavior (Doyle et al., 2019; Lunsford & Oviatt, 2006). However, no research in our corpus attempts to explore whether this phenomenon really happens more massively in chatbot text-based interaction than in speech-based communication.

To theoretically account for this and other supposed peculiarities of chatbot technology, researchers could refer to theories that focus on mediated conversations specifically conducted through *written language*, like certain CMC theories (e.g., Joinson, 2001; Tidwell & Walther, 2002; Valkenburg & Peter, 2009). CMC research is used by only four papers in our corpus (Beattie et al., 2020; Sundar et al., 2016; Hill et al., 2015; Lee et al, 2020a), but could give relevant insights for studying chatbots. Going back to our example on humanness, the hyperpersonal theory (Walther, 1996) highlights that text-based CMC is limited in the amount of information conveyed. As a consequence, interlocutors are likely to fill in the gaps by using their own mental schemas or supposed known information about the sender, like shared category membership (Spears & Lea, 1992), confirming their hopes and positive expectations (Finkel et al., 2012). We may thus ask ourselves: Are these projections of humanness favored by the 'incomplete' nature of text-based interaction? Can we explain such projections by using CMC theories, or do we need to develop new specific theories when one of the actors is a machine? A promising line of future research is to give answers to these and other research questions investigating the distinctive traits of interaction with text-based chatbots.

5.2 Using well-defined, grounded, and shared constructs

A variety of articles we reviewed investigated specific theoretical constructs relevant for the interaction, such as social presence (Araujo, 2018; Go & Sundar, 2019) and self-disclosure (Lee et al., 2020b). Often, this kind of works selectively use constructs from one or more existing theories, rather than employing an entire conceptual framework. For instance, Lee et al. (2020b) focus on the construct of self-disclosure as understood in social penetration theory, which explains how interpersonal relationships unfold, but do not consider in their design the stages of self-disclosure, which describe how self-disclosure evolves during interaction, or the barriers of self-disclosure, which can slow down the process of opening up (Altman & Taylor, 1973). This likely happens because single constructs translate more easily into design features: however, it might also lead to design chatbots based on constructs that do not work independently but only when paired with other constructs, making designers miss an important part of the picture (Hekler et al., 2013).

However, the reviewed research also uses concepts and constructs that are not grounded on specific theories, which may lead to additional problems. For instance, as we have seen, a variety of papers investigate 'emotional' chatbots. These studies seem to carelessly treat what emotions mean, simply differentiating chatbots that are able to understand and express emotions from non-emotional chatbots. For example, Kuramoto et al. (2018) based the design of a conversational agent on a theory for changing human attitude by conversation (i.e., Balance Theory by Heider, 1946) in order to reduce users' anger during the interaction. However, the authors do not refer to any theory or study explaining what anger is and why it is experienced. In fact, anger appears to be provoked by different triggers, such as other people, psychological and physical distress, and intrapersonal demands (Kashdan et al., 2016), and different individuals may experience anger in different ways and for different reasons (e.g., Jones et al., 2011; Pease et al., 2015). All these factors are not accounted in the user study reported in the article. This might produce a partial picture of how users truly react to the chatbot.

This said, the surveyed articles often also disagree on the exact definitions of the constructs that they utilize. For instance, research investigating trust either does not define trust, somehow relying on 'common sense' (Jin et al., 2019; Zamora, 2017), or proposes new explanations of trust based on novel findings (Følstad et al., 2018), or refers to already established definitions of trust: these definitions often come from different disciplines and operationalize the construct differently, which may lead to different 'measures'. For instance, Zhou et al. combine three trust measures (integrity, competence, and benevolence) based on a model developed by Mayer et al. (1995) to create an index measure called "trust"; while Chattaraman et al. (2019) assessed trust using an 8-item trust scale defined by Klein (2007). As a result, it becomes difficult to compare the outcomes of different studies. A lack of consistency in the concepts being measured across different studies has been also found by Clark et al. (2019a) with reference to speech-based interaction: the authors highlighted difficulties for the domain in building robust measures and a body of knowledge around specific concepts.

Text-based chatbot research would thus benefit from greater consensus on the employed constructs: researchers should better define and ground them in theories considered relevant by the research

community. This effort may eventually lead to a shared 'vocabulary' mapping how specific constructs have been used in chatbot research and what definitions and measures are most useful for the understanding of the human-chatbot interaction experience.

5.3 Building on previous work to further existing theory

The review makes apparent that the field lacks unified models and theories that may give explanation of fundamental aspects of the interaction experience with chatbots. Reviewed research frequently leads to contrasting or scattered results, and even though these differences are retraced to individual and contextual factors (included the goal of the chatbot), the authors often miss to meaningfully build on previous work. This may be also due to the lack of consensus on fundamental constructs that are relevant for the chatbot domain, which we have pointed out in the previous sub-section. For instance, many reviewed papers stress the role of expectations in shaping the user's perceptions of the chatbot during the interaction (Schumaker & Chen, 2010; Akhtar et al., 2019; Følstad & Skjuve, 2019; Jain et al., 2018; Zamora, 2017), even influencing her level of engagement (Portela and Granell-Canut, 2018), reactions to the chatbot's emotional expressions (Urakami et al., 2019), and attitudes toward the chatbot itself (Go & Sundar, 2019). However, no research attempts to categorize the different kinds of expectations that users may have, in which contexts they are more likely to be activated, how they are formed, and what diversified effects they may produce on the interaction. This shows how the research within the chatbot domain is fragmented, emphasizing a need for connecting findings and providing order across different studies. Future research needs to better build on previous work and contribute to develop coherence and cohesion in the current body of knowledge. The current lack of papers attempting to provide unified theoretical frameworks for human-chatbot interaction signals that there is room for HCI to develop novel theories of human-chatbot interaction.

5.4 Widening the notion of 'context'

The papers reviewed in this article reveal a tendency to conduct laboratory experiments to study text-based chatbot interaction. This approach often employs WoZ systems or controlled prototypes. Even though this research is paramount for developing knowledge on chatbot interaction, more effort is needed to understand how people interact with this technology in real situations of use and how findings gained in lab may transfer to real world. Nonetheless, albeit not predominant, studies 'in the wild' are present in our corpus. This research analyzes either the interaction of users with already existing chatbots (e.g., Li et al., 2020), or develop new prototypes and deploy them in 'real-world contexts' (e.g., Seering et al., 2020; Liao et al., 2018).

What is worth to notice here, nonetheless, is the narrow notion of context that even these studies embrace, as well as the lack of research exploring wider implications of chatbot use. Field studies on chatbots in our corpus take place in e.g., organizational (Cranshaw et al., 2017; Kimani et al., 2019), physical (Chaix et al., 2019) and mental (Fitzpatrick et al., 2017) health, customer support (Li et al., 2020) and entertainment (Seering et al., 2020) contexts. In these works, the horizon in which the technology is studied is limited to the context of immediate use of the technology.

For instance, Calendar.help (Cranshaw et al., 2017) is evaluated mainly with reference to the improvements in productivity that it can bring in a multinational technology corporation. A chatbot adopted within an organization, however, commonly does not impact only on workers who use it: rather, it may change organizational processes and structures, affect customer experience and other partner organizations, and so on. Calendar.help's architecture, in fact, involves a human-in-the-loop mechanism, in which humans intervene when the chatbot is not able to handle the user's requests. There are no mentions in the article, however, that these latter workers could be negatively affected by the chatbot, being used as humbots, i.e., humans that do bots' work (Grudin & Jacques, 2019). If users are not aware that these chatbots are humans (like Calendar.help's users), they could use them as AI servants, requiring services around the clock and somehow constraining those workers to align to machine's rhythms. This might lead to negatively organizational changes, increasing inequalities

among employees and incrementing the number of minimum-wage workers in charge of executing machine-like tasks.

Likewise, chatbots in the health domain may have wider implications than the immediate benefits brought to the user. Beyond the effectiveness of the target intervention (e.g., alleviating symptoms of depression and anxiety like in Fitzpatrick et al., 2017), deploying a health chatbot may involve issues as diverse as the role of the human doctor and how the chatbot mediates or substitutes the relationship of the patient with her, who will have access to and handle the health data collected by the chatbot, how such data should be interpreted, how the chatbot will be integrated into a certain healthcare system and so on. At present, such issues appear to be mostly ignored by literature and by the ongoing debate about chatbots. However, uncertainty about e.g., privacy and 'substitution of humans' issues may play a relevant role in preventing users from using chatbot technology (De Barcelos Silva et al., 2020; Völkel et al., 2020; Følstad et al., 2018; Portela and Granell-Canut, 2018; Nadarzynski et al., 2019).

This is to say that studies on chatbots need to be situated in wider horizons than the limited context in which the technology is deployed. We have to notice that a narrow perspective on context is quite common in technology research. For instance, it has been emphasized that technologies aimed at changing behaviors focus on the target behavior to be changed, while ignoring the other aspects of the user's life which may be relevant for the process of change (Rapp et al., 2019). Considering a wider notion of context is even more important when technology is an agent, which can autonomously undertake actions, behaving as and even substituting a real person, because of the social and organizational impacts it may have. A research challenge for HCI that should be tackled in the future, therefore, is to account for such implications and even side-effects, encouraging us to rethink our notion of 'context of use' of chatbot technology.

5.5 Promoting design research

This review shows that there is a considerable interest in evaluating the chatbot design with reference to its user experience. However, it also makes evident that literature lacks design related research. There are no papers that seriously tackle design issues. This absence parallels the lack of design studies in speech interface domain (Clark et al., 2019a). Methods traditionally employed in interaction design, such as participatory design (Schuler & Namioka, 1993), cultural probes (DeHaemer & Wallace, 1992), and research through design (Zimmerman & Forlizzi, 2008) are not even mentioned in our corpus. Even though several papers formulate some 'implications for design' as a consequence of user studies on specific chatbots (e.g., Jain et al., 2018; Lee et al., 2020b; Xiao et al., 2020; Li et al., 2020; Chattaraman et al., 2019; Candello et al., 2019; Lee et al., 2019; Liao et al., 2018), the lack of design research shows that there is still not a clear understanding of what are the more pressing design challenges that need to be solved in the chatbot domain. The fragmented nature of chatbot research is apparent even with reference to design issues, since the 'implications for design' provided in the articles are rarely discussed against previous design guidelines. As a result, there is no wide consensus on grounded strategies on which build novel design work. With the chatbot tsunami, nonetheless, the need for design knowledge is becoming more pressing. In particular, research through design, which applies design practice to explore problems in order to create new knowledge, even through the creation of 'critical' (Bardzell & Bardzell, 2013) or 'fictional prototypes' (Blythe et al., 2014), could bring interesting insights for the future evolution of chatbots[1]. This said, HCI research should find those design research techniques that could work better with chatbots, a technology that mainly relies on written texts and thus requires the design of the 'language itself'.

---

[1] Critical design is a research through design methodology that foregrounds the ethics of design practice and explores alternative design values through the creation of provocative prototypes (Bardzell & Bardzell, 2013). Design fictions are fictional prototypes that do not exist but can be used to explore the long-term and systemic consequences of a certain technology (Rapp, 2019, 2020).

5.6 Be human or machine? Tackling ethical issues

Previous research on voice-based intelligent personal assistants found that incongruences between expectations about humanness and real interaction may hinder the user experience (Luger & Sellen, 2016; Leahu et al., 2013). As we have seen, humanness can lead to contrasting outcomes in text-based chatbots as well, seemingly depending on the user's beliefs on how a chatbot should be (Liu & Sundar, 2018), how the ongoing conversation is perceived against her expectations (Go & Sundar, 2019) and the context in which the chatbot is used (Svenningsson & Faraon, 2019). For many users, however, perceiving the chatbot as endowed with human-like aspects (like empathy, and self-disclosure) improves their conversational experience, primarily encouraging them to feel more 'favorable' (in terms of trust, openness, tolerance, etc.) toward the chatbot itself.

Given these positive outcomes, it comes as no surprise that much research we reviewed is struggling to design chatbots endowed with human-like qualities. Designers may rely on the users' tendency in projecting humanity into the chatbot (e.g., Araujo, 2018; Go & Sundar, 2019; Beattie et al., 2020) to make them 'believe' that they are interacting with a human. However, it is quite remarkable that a critical debate tackling the ethical issues arising from this 'subtle deception', be either intentionally cueing perceptions of humanness, or blurring the distinction between humans and machines in human-in-the-loop systems, or simply letting the user free to believe what she wants, is completely absent in our corpus.

Discussing chatbot design only with reference to its effectiveness, usefulness, and capability of satisfying and engaging people seems not to be sufficient. There could be cases in which making the user experience more engaging might not be a valuable outcome or encouraging a favorable disposition toward the chatbot might be questionable. Some scenarios may clarify this point.

For example, a chatbot designed to collect health data may exploit its human-like qualities in order to gain information that a user may not have otherwise disclosed to a machine. In turn, this data may be mined and transmitted to other machines, further confounding this problem. Likewise, a user could easily become emotionally attached to an artificial agent only because it is endowed with cues that

continuously activate cognitive heuristic about characteristics of humanness and not because she purposefully attempts to develop a human-like relationship with an artificial entity. In the same vein, the user could recruit artificial alter egos in charge of maintaining relationship with friends and significant others, making them believe that they are interacting with a human, while actually their interlocutor is a chatbot.

In sum, what are the impacts that a pervasive employment of human-like chatbots design could produce on interpersonal relationships? Are there domains in which human-like chatbots should not be used? What if novel, effective human-like chatbots will be successfully deployed and become so pervasive that most of our computer-mediated social interactions will be with an artificial partner? These questions did not receive an answer in the articles reviewed here, and we think that HCI researchers should start investigating them in the next years.

5.7 Limitations

The goal of this review was to identify the main topics discussed across human-chatbot interaction research, focusing on the human side of such interaction. This entailed the exclusion of works primarily focused on technology. Moreover, we decided to narrow our analysis to papers published in journals and main conference proceedings, ignoring research appeared in workshops, posters and adjunct proceedings. Even though these venues go beyond the scope of this review, similar themes are present in works published in there. Finally, we chose to pay attention to studies focusing on chatbot communicating through written language. Future reviews could go in depth in comparing our findings with those pointed out by other reviews that focused on speech interfaces and embodied conversational agents.

6 CONCLUSION

This literature review points to five main research themes that are recurrent across the 83 papers we analyzed and identifies the main methodological approaches adopted in research. It highlights that researchers explore users' experiences, attitudes and behaviors towards chatbots through a variety of commercial, WoZ, and fully developed prototype systems, primarily using an experimental approach. The review also points out that research particularly focused on the exploration of whether and why users accept the chatbot technology and how trust, engagement and satisfaction may be relevant aspects of their interaction experience. Then, we emphasize research investigating how conversational issues may hinder the progress of the interaction and the role that emotions may play during the conversations with chatbots. Finally, we notice that reviewed research paid particular attention to how people ascribe humanness to chatbots, also identifying the effects that humanness produces on the interaction experience and highlighting papers proposing prototypes in which humanness is a pivotal design element. Based on these findings, we identified several lines of future research that could be interesting to explore for the HCI community, stressing that scholars should develop theory addressing the peculiarity of human-chatbot interaction, use well-defined and shared constructs, better build on previous work, widen the notion of context when exploring the implications of technology, conduct research focused on design, and tackle relevant ethical issues mainly arising from the users' tendency to ascribe humanness to chatbots.

Table 7. Articles included in the review.

Article	Goal	Sample	Method	Main findings
Akhtar et al. (2019)	To analyze whether human-chatbot interactions are informative in determining users' topic of interests and user satisfaction.	215,859 different sessions of human-chatbot interaction	Quantitative, conversation log analysis.	The majority of users showed a low degree of satisfaction: they expected to receive an answer to the stated question quickly, otherwise they left the conversation soon.
Araujo (2018)	To explore the extent to which human-like cues can influence perceptions about social presence and anthropomorphism.	175 participants	Quantitative, between-subjects experiment, questionnaire.	The usage of human-like language or name are sufficient to increase perception of the agent as being human-like.
Ashktorab et al. (2019)	To investigate user's repair strategies for conversational breakdowns with chatbots.	203 participants	Mixed-methods, pairwise comparison experiment, survey.	Participants preferred strategies providing options and explanations, as they manifested chatbot's initiative.
Avula et al. (2018)	To investigate the effects of a searchbot on participants' collaborative experience and engagement.	27 pairs of participants (n=54)	Quantitative, Wizard of Oz within-subjects experiment, questionnaire.	The searchbot improved participants' collaborative experience. Time of intervention is the key for engagement
Beattie et al. (2020)	To analyze the effects of emojis on the user's perception of a chatbot.	96 participants	Quantitative, between-subjects experiment, questionnaire.	Participants prefer emoji-featuring messages and evaluated chatbots using emojis similarly to humans.
Beriault-Poirier et al. (2019)	To compare user experience between chatbots and websites.	10 participants	Mixed methods, usability test, interviews, observation, questionnaire.	Users preferred using websites rather than chatbots, which had a higher abandon rate.
Brandtzaeg & Følstad (2017)	To investigate users' motivation for using chatbots.	146 participants	Mixed-methods, survey with questionnaire and open-ended answers.	The most frequently reported motivational factor is "productivity", as chatbots help users obtain efficient assistance or information.
Candello et al. (2017)	To study the influence of chatbot's visual aspects on the perception of humanness.	199 participants (90+109)	Mixed-methods, Turing Test-inspired between-subjects experiment, questionnaire, observational notes.	Machine-like typefaces are judged by humans as "robotic", whilst handwritten-like typefaces are not perceived as more human.
Candello et al. (2019)	To investigate the effect of a social audience on human-chatbot interaction.	More than 5,000 conversation logs and video recordings	Mixed-methods, self-report questionnaire, semi-structured interviews and conversation log analysis.	The presence of audience has a significant effect on interaction with chatbots in public spaces.
Cervone et al (2018)	To investigate the feasibility of learning evaluation models only relying on conversation-level human ratings.	4,967 human-chatbot conversations	Quantitative, conversation log analysis, users' rating.	The length of conversation appeared to be positively correlated with user enjoyment.

Chaix et al. (2019)	To analyze the user experience of patients with breast cancer while conversing with a chatbot.	4737 (conversation log analysis) + 958 participants (survey)	Mixed-methods, conversation log analysis, survey, focus group.	The experience with the chatbot was appreciated and increased patients' compliance with the treatment.
Chattaraman et al. (2019)	To investigate task-oriented and social-oriented chatbots' benefits on older users in an online shopping task.	121 participants	Quantitative, between-subjects experiment, questionnaire.	Users with low Internet experience prefer using task-oriented chatbots, whilst users with high Internet experience find more trustworthy the social-oriented chatbot.
Chung et al. (2018)	To analyze user satisfaction with a luxury brand e-commerce chatbot.	161 participants	Quantitative, survey.	Perceived communication accuracy and credibility positively influence satisfaction
Ciechanowski et al. (2017)	To investigate if chatbots with and without a human-like avatar image may engender uncanny feelings.	31 participants in two phases of the research (quantitative and qualitative)	Mixed-methods, between-subjects experiment, psychophysiological signals, questionnaire, debriefing sessions.	Users prefer the text-based chatbot, rather than the avatar-embedded chatbot, which engendered the uncanny valley effect.
Corti & Gillespie (2016)	To investigate the differences in initiating repairs of misunderstandings with non-human and human-like chatbot.	108 participants	Quantitative, between-subjects experiment, conversation log analysis.	Users prefer to repair the misunderstandings with human-like agents, rather than with text-based chatbots.
Cranshaw et al. (2017)	To analyze and evaluate a scheduling conversational agent.	11 (study 1) + 24 (study 2) + 178 participants (study 3)	Mixed-methods, Wizard of Oz study, usability test, field study, interaction log analysis, email analysis, interviews.	The chatbot provided a good user experience and high engagement by helping people in their productivity, and it was often perceived as humanlike.
de Kleijn et al. (2019)	To investigate the linguistic cues that users use to understand whether their interlocutor is a human or a chatbot.	53 participants	Quantitative, Turing Test-inspired experiment, within-subjects (first task) and between-subjects (second and third tasks) experiment, ratings.	Center-embedded sentences are perceived as less humanlike than right-branching sentences.
de Medeiros et al. (2019)	To analyze and evaluate a tutoring conversational agent capable of enunciating "small talk" segments.	48 participants, 1732 phrases, 491 excerpts of small talk	Mixed-methods, survey, open-ended questions.	The small talk feature plays a role in making the tool more approachable and has beneficial effects on the students' performances.
Denecke et al. (2018)	To evaluate the user experience of a self-anamnesis chatbot prototype.	22 participants	Mixed-methods, usability test, questionnaire, qualitative feedback.	The chatbot is a practical way for collecting anamnesis data. Users felt engaged in answering its questions.
Denecke et al. (2020)	To evaluate the user experience of a chatbot addressed to patients with emotion regulation problems.	21 participants	Mixed-methods, usability test, questionnaire, qualitative feedback.	The chatbot is able to encourage patients elaborate on their emotions
Fadhil et al. (2018)	To analyze the effects of different chatbots' dialogue styles.	58 participants	Quantitative, mixed design experiment, questionnaire.	Emojis-enriched texts are preferred in mental health situations, whilst plain-text is preferred in physical health contexts.

Fiore et al. (2019)	To study the user acceptance and experience of a task-oriented chatbot.	12 participants	Quantitative, usability test, questionnaire.	Participants were positive about the chatbot prototype and appreciated its simplicity, and pro-active guidance.
Fitzpatrick et al. (2017)	To determine the effects of a chatbot based on cognitive-behavioral therapy on users' symptomatology.	70 participants	Mixed-methods, between-subjects experiment, questionnaire, open-ended questions.	Participants significantly reduced their symptoms of depression while using the chatbot.
Følstad & Skjuve (2019)	To investigate users' perceptions and attitudes toward chatbots for customer service.	24 participants	Qualitative, semi-structured interviews.	Participants had realistic expectations about chatbots and human likeness of chatbots for customer service has minor importance.
Følstad et al. (2018)	To investigate the factors affecting trust toward chatbots for customer service.	13 participants	Qualitative, semi-structured interviews.	Chatbots' ability to correctly interpret the users' requests and to provide helpful responses are key factors that impact on trust.
Galko et al. (2018)	To evaluate a chatbot for submitting university applications.	5 participants	Mixed-methods, usability test, post-session rating, time on task, error rate.	Participants identified several issues related to the chatbot (informality, inconsistency of options, negative language).
Go & Sundar (2019)	To analyze the effect of anthropomorphic cues on user engagement.	141 participants	Quantitative, between-subjects experiment, questionnaire.	A high level of conversation contingency compensates for low anthropomorphic visual cues.
Greer et al. (2019)	To examine the effects of a healthcare chatbot on key well-being outcomes in adults treated for cancer.	45 participants	Mixed-methods, between-subjects experiment, questionnaire, open-ended questions.	The chatbot provides a useful and acceptable format for young adults to connect with a positive psychology intervention.
Hill et al. (2015)	To analyze the differences between human-human and human-chatbot communication.	100+100 conversations	Quantitative, conversation analysis.	There are notable differences between human-human and human-chatbot communication in the content and quality of the conversations.
Holmes et al. (2019)	To evaluate the usability of a healthcare chatbot.	26 participants	Mixed-methods, usability test, questionnaire, observation, thinking aloud protocol.	Conventional methods for assessing usability and user experience are not sufficient and accurate when applied to chatbots.
Huang et al. (2019)	To analyze acceptance and resistance towards AI in customer service.	411 participants	Quantitative, survey.	Perceived value affects behavioral intention to accept chatbots for customer care.
Jain et al. (2018)	To understand how the actual perceptions of chatbots' users are shaped by their expectations.	16 participants, 10,000 messages human-chatbot.	Mixed-methods, conversation log analysis, semi-structured interviews.	Participants' expectations were not met by the chatbots, which turned into a very unsatisfactory experience.
Jin et al. (2019)	To evaluate a chatbot for music recommendations.	45 participants	Quantitative, within-subjects experiment, questionnaire, interaction log analysis.	Musical sophistication seems to increase users' acceptance and trust toward the chatbot.

Kim et al. (2020)	To evaluate of a chatbot that facilitates group discussion.	65 (study 1) 25 (study 2) + 20 (study 3)	Mixed-methods, survey, mixed design experiment, focus group, semi-structured interviews, questionnaire, conversation log analysis.	The chatbot served as a group member, promoting members' participation and discussion.
Kimani et al. (2019)	To evaluate a chatbot that promotes productivity and well-being at work.	29 participants	Mixed-methods, field study, questionnaire, interaction log analysis, experience sampling responses, qualitative feedback.	The chatbot improved productivity at workplace and participants were more mindful about their work practices.
Kuramoto et al. (2018)	To evaluate a chatbot aimed at regulating workers' anger in a customer support working environment.	18 participants	Mixed-methods, within-subjects experiment, questionnaire, interviews	Participants' anger can be suppressed by using the chatbot, but the anger itself is at a higher level when the agent is present.
Lee et al. (2019)	To examine how people can care for another being (chatbot) as a way to care for themselves.	12 (pilot) + 67 (experiment) participants	Mixed-methods, between-subjects experiment, questionnaire, open-ended responses.	Participants treated the chatbot as a human social agent and their self-compassion increased when they were compassionate towards the chatbot.
Lee et al. (2020a)	To investigate whether users' perception of a mind within a chatbot is associated with feelings of co-presence, closeness and intention to use.	64 participants	Quantitative, Wizard of Oz between-subjects experiment, questionnaire.	User's perception of a mind within a chatbot leads to stronger sense of co-presence and closeness.
Lee et al (2020b)	To evaluate a chatbot that has self-disclosure features when it performs small talk with people.	47 participants	Mixed-methods, between-subjects experiment, conversation log analysis, semi-structured interviews, questionnaire	Computer agents' self-disclosure may facilitate the users' self-disclosure and have positive impact on perceived intimacy and enjoyment with the chatbot.
Li et al. (2020)	To analyze non-progressing conversations and users' coping strategies to deal with them.	19,451 exchanges involving 1,837 users	Qualitative, conversation log analysis.	Non-progresses mainly occurred when the users requested information and the chatbot did not recognized users' intent.
Liao et al. (2018)	To analyze the conversations that users have in the wild with a chatbot.	115 (survey) + 337 participants, 6004 messages analyzed	Quantitative, field study, survey, conversation log analysis.	Chitchat, frequent occurrences of second-person pronouns, and casual texting are positive signals of playful satisfaction.
Liu & Dong (2019)	To evaluate the user experience of a question answering chatbot.	30 participants	Mixed-methods, usability test, questionnaire, conversation log analysis, observation.	Users satisfaction increased using the chatbot in route navigation task.
Liu & Sundar (2018)	To investigate whether a chatbot for advice about a personal problem should refrain from offering emotional support.	158 (study 1) + 88 (study 2)	Quantitative, between-subjects experiment, questionnaire.	Expression of sympathy and affective empathy is perceived as more supportive than advice-only.

Ly et al. (2017)	To investigate the effectiveness of a fully automated self-help intervention.	28 participants	Mixed-methods, mixed design experiment, questionnaire, semi-structured interviews (sub-sample of 9 participants).	Participants show high engagement during the 2-week long intervention. The chatbot is perceived somewhat as a real person.
Medeiros et al. (2019)	To evaluate a chatbot designed to give emotional support.	6 participants	Qualitative, focus group.	Participants generally appreciated the chatbot, but they reported limitations.
Mendez et al. (2019)	To explore explores the potential use of chatbots in future faculty mentoring	10 participants	Qualitative, focus group	Chatbots may be effectively employed for mentoring for students, yet a more sophisticated chatbots are needed.
Michaud (2018)	To analyze users' interaction with a healthcare chatbot.	1,023 texts sent by 491 distinct accounts	Mixed-methods, conversation log analysis.	A chatbot should be able to respond to "purely dialogue" interactions.
Mori et al. (2019)	To study how humans identify human-likeness in Q&A online communication.	30 participants	Quantitative, Turing Test-inspired experiment, questionnaire.	The chatbot-created texts left negative impressions (e.g. questions and answers don't match up).
Morrissey & Kirakowski (2013)	To identify dimensions on which users might judge the naturalness of a chatbot.	14 (study 1) + 20 (study 2) participants	Mixed-methods, Critical Incident Technique experiment, open ended questions, questionnaire.	Users judge the naturalness of a chatbot by relying on conscientiousness, originality, manner, and thoroughness.
Mou & Xu (2017)	To investigate whether humans would reveal their personality traits differently in interacting with a chatbot and a human.	245 participants	Quantitative, within-subjects experiment, questionnaire.	Users were more open, agreeable, extroverted, conscientious and self-disclosing when they interacted with humans than with a chatbot.
Nadarzynski et al. (2019)	To explore participants' willingness to engage with health chatbots.	216 (survey) + 29 (interviews) participants	Mixed-methods, survey, semi-structured interviews.	The curiosity about new health technologies was a strong predictor of acceptability.
Park et al. (2018)	To evaluate how well two chatbots emulate human conversation.	58 participants	Quantitative, questionnaire, conversation log analysis.	Participants used fairly simple language and made a variety of spelling and grammatical errors during interaction.
Pérez-Marín & Pascual-Nieto (2013)	To investigate the effect of a pedagogic conversational agent (PCA) on students.	22 participants	Mixed methods, within-subjects experiment, observation, questionnaire, interviews.	Children perceived the chatbot as a human and a friend. This believed humanness led them to exhibit prosocial behaviors.
Portela & Granell-Canut (2017)	To understand the possibilities of engagement and affection in the use of chatbots	13 participants	Mixed-methods, Wizard of Oz, usability test, questionnaire, semi-structured interviews.	Users were strongly engaged when the chatbot showed empathic behaviors, which provoked the same empathic feeling.
Procter et al. (2018)	To improve the communication between students and chatbot-based intelligent tutoring systems.	56 participants	Quantitative, survey, conversation log analysis.	The quality of agent-based interventions influenced whether students find the chatbot useful.

Ruan et al. (2019)	To evaluate the effects on learning of a chatbot and to compare it with a traditional learning system.	47 (study 1) + 76 (study 2)	Mixed methods, within-subjects experiments, questionnaire, conversation log analysis, qualitative feedback.	Participants were more engaged in learning with the chatbot than with traditional instruments, but did not spend all the time in learning.
Sanny et al. (2020)	To analyze the customer satisfaction factors that influence chatbot acceptance in Indonesia.	119 participants	Quantitative, survey.	Usefulness, brand image, personality, and ease of use influence user acceptance.
Schuetzler et al. (2019)	To analyze the effect of the ability of the chatbot to mimic human conversations on behavioral indicators of deception.	103 participants	Quantitative, deception experiment, conversation log analysis.	Increased conversational skills encourage strategic deception behavior.
Schumaker & Chen (2010)	To analyze the quality of interaction and user satisfaction of six ALICE chatbots.	347 participants, 33,446 interactions	Quantitative, conversation log analysis.	The responses to interrogatives beginning with "Are" and "Where" resulted in higher response satisfaction levels.
Seering et al. (2020)	To analyze human-chatbot interaction with a chatbot that grows up and evolves in an online community.	46 participants, 5716 messages	Qualitative, field study, conversation log analysis.	BabyBot was able to interact with everyone in the community. User interaction changed as the chatbot grew up.
Sensuse et al. (2019)	To evaluate the benefits of a customer service chatbot in a company.	43 participants	Quantitative, survey.	Information quality, service quality and intention to use influenced users' satisfaction.
Shi et al. (2020)	To analyze the effect of perceived identity and inquiry strategies of the chatbot on persuasion.	790 participants	Quantitative, between-subjects experiment, questionnaire.	Perceived identity of the chatbot had significant effects on the persuasion outcome and interpersonal perceptions.
Skjuve et al. (2019)	To analyze the uncanny valley effect of a text-based chatbot.	28 participants	Mixed-methods, between-subjects experiment, questionnaire, qualitative feedback.	The lack of transparency does not lead to uncanny valley effect, even when there is uncertainty related to chatbot's true nature.
Smestad & Volden (2018)	To investigate the impact of personality matching between chatbots and users.	16 participants	Quantitative, between-subjects experiment, questionnaire.	Personality has significant positive impact on the user experience of chatbots
Sundar et al. (2016)	To investigate the relationships between contingency, message interactivity and engagement.	110 participants	Quantitative, between-subjects experiment, questionnaire.	The presence of a chatbot in a website boosts the level of message interactivity, which may lead to engagement
Svenningson & Faraon (2019)	To analyze the factors that are related to the perception of humanness.	69 participants	Quantitative, survey.	Conversations with human-like chatbots are not always welcomed by users.
Ta et al. (2020)	To investigate the user experience of a chatbot designed to provide social support.	4434 reviews (study 1) + 66 participants (study 2)	Qualitative, qualitative text analysis, survey with open-ended questions.	The chatbot appeared to be a promising source of companionship, providing emotional support.

Tärning & Silvervarg (2019)	To analyze the effects of a pedagogical agent on learning outcomes.	89 participants.	Quantitative, between-subjects experiment, questionnaire, conversation and interaction log analysis.	it is more beneficial to design a digital tutee with low self-efficacy than one with high self-efficacy.
Trivedi (2019)	To assess the customer experience of using banking chatbots.	258 participants	Quantitative, survey.	Information quality, system quality and service quality have a significant impact on customer experience of banking chatbots.
Urakami et al. (2019)	To examine user's perceptions of expressions of empathy by a chatbot.	150 participants	Quantitative, survey.	Statements expressing help or showing interest are evaluated as highest in regard to empathy by participants
Van den Broeck et al. (2019)	To analyze whether and how perceived helpfulness and usefulness of a chatbot affects perceived instrusiveness of advertising in a later stage.	245 participants	Quantitative, survey, conversation log analysis.	Perceived intrusiveness of chatbot advertising was found to be dependent on perceived helpfulness and perceived usefulness of the chatbot
Völkel et al. (2020)	To study the users' attitudes towards chatbots able to automatically assess users' personality during interaction.	21 participants	Mixed-methods, within-subjects experiment, questionnaire, interviews.	Participants regarded personality as very sensitive data.
Wang & Nakatsu (2013)	To analyze the interaction with a conversational agent which mimics Confucius.	1029 interactions (115 users)	Qualitative, conversation log analysis.	Bad emotions, like unhappiness or frustration, may be elicited by irrelevant responses provided by the chatbot.
Warwick & Shah (2015)	To analyze how humans identify machines in a Turing test.	30 participants	Qualitative, Turing Test, conversation log analysis.	Participants based their own judgments on the interlocutor's capabilities of lying and being humorous.
Westerman et al. (2019)	To analyze the perceptions of humanness of chatbots.	83 participants	Quantitative, Wizard-of-Oz between-subjects experiment, questionnaire, conversation log analysis.	Capitalized words do not influence perceived humanness, whereas typos make the chatbot much less human-like.
Xiao et al. (2020)	To evaluate the user experience and effectiveness of an interview chatbot with active listening skills.	206 participants	Quantitative, conversation log analysis, between-subjects experiment, questionnaire.	Chatbots with active listening skills are more effective at engaging users and eliciting quality user responses.
Xu et al. (2017)	To evaluate a chatbot designed to automatically generate responses for users' requests on social media.	600 participants (survey) + 200 requests	Mixed-methods, survey, qualitative conversation log analysis.	Over 40% of the users' requests were not informational but emotional.
Yen and Chiang (2020)	To investigate factors that influence customers' trust in chatbots.	204 (study 1) + 30 (study 2) participants	Quantitative, survey, EEG within-subjects experiment.	Credibility, competence, anthropomorphism, social presence, and informativness influence customers' trust.

Zamora (2017)	To understand user's perceptions and expectations of chatbots.	54 participants	Mixed methods, survey (with also open-ended questions), interaction log analysis.	Participants had undefined expectations. They compared the experience to services with which they were familiar.
Zarouali et al. (2018)	To identify the determinants that influence the likelihood to use and recommend the chatbot.	245 participants	Quantitative, survey.	Attitude toward the brand influences the likelihood to use and recommend the chatbot.
Zaroukian et al. (2017)	To investigate automation bias for confirming erroneous information with a chatbot.	161 participants, 2482 reports	Quantitative, between-subjects experiment, task performance measures.	Automation bias occurs when users trust the agent too much.
Zhou et al. (2019)	To investigate how the personality of a chatbot interviewer and user's personality influence the user's trust in the chatbot.	1280 participants	Quantitative, field study, users' ratings.	Users' personality traits influence users' perception of and their willingness to confide in and listen to a chatbot interviewer
Zumstein & Hundertmark (2017)	To analyze chatbot acceptance in a public transportation company.	134 participants	Quantitative, survey.	Users showed high acceptance of the chatbot, albeit only 40% said that they would reuse it.

REFERENCES

Abdelkefi, M., & Kallel, I. (2016). Conversational agent for mobile-learning: A review and a proposal of a multilanguage text-to-speech agent, "MobiSpeech". In *Proceedings of the 2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS)* (pp. 1-6), IEEE. doi: 10.1109/RCIS.2016.7549294

Akhtar, M., Neidhardt, J., & Werthner, H. (2019). The potential of chatbots: analysis of chatbot conversations. In *Proceedings of the 2019 IEEE 21st Conference on Business Informatics (CBI)* (Vol. 1, pp. 397-404), IEEE. doi: 10.1109/CBI.2019.00052

Altman, I., & Taylor, D. A. (1973). *Social penetration: The development of interpersonal relationships*. Holt, Rinehart & Winston.

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior, 85*, 183-189. doi: 10.1016/j.chb.2018.03.051

Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient chatbots: repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1-12, ACM, New York. doi: 10.1145/3290605.3300484

Attfield, S., Kazai, G., Lalmas, M. & Piwowarski, B. (2011). Towards a science of user engagement (position paper). In *Proceedings of the WSDM Workshop on User Modeling for Web Applications*, 9-12.

Avula, S., Chadwick, G., Arguello, J., & Capra, R. (2018). Searchbots: User engagement with chatbots during collaborative search. In *Proceedings of the 2018 conference on human information interaction & retrieval*, 52-61, ACM, New York. doi: 10.1145/3176349.3176380

Bardzell, J. & Bardzell, S. (2013). What is "critical" about critical design? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, 3297–3306. doi: 10.1145/2470654.2466451

Beattie, A., Edwards, A. P., & Edwards, C. (2020). A Bot and a Smile: Interpersonal Impressions of Chatbots and Humans Using Emoji in Computer-mediated Communication. *Communication Studies*, 71(3), 409-427. doi: 10.1080/10510974.2020.1725082.

Beriault-Poirier, A., Tep, S. P., & Sénécal, S. (2018). Putting Chatbots to the Test: Does the User Experience Score Higher with Chatbots Than Websites?. In *Proceedings of the International Conference on Human Systems Engineering and Design: Future Trends and Applications*, 204-212. Springer, Cham. doi: 10.1007/978-3-030-02053-8_32

Blythe, M. (2014). Research through design fiction: narrative in real and imaginary abstracts. In *Proceedings of the SIGCHI conference on human factors in computing systems*, ACM, New York, 703-712. doi: 10.1145/2556288.2557098

Boiteux, M. (2018). Messenger at F8 2018. *Medium*. Blog post retrieved from https://blog.messengerdevelopers.com/messenger-at-f8-2018-44010dc9d2ea. Accessed September 20, 2019.

Boyle, E. A., Connolly, T. M., Hainey, T., and Boyle, J. M. (2012). Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior*, 28(3), 771-780. doi: 10.1016/j.chb.2011.11.020

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. In *International Conference on Internet Science*, 377-392. Springer, Cham. doi: 10.1007/978-3-319-70284-1_30

Brave, S. & Nass, C. (2002) Emotion in human-computer interaction, in (eds. J. A. Jacko & A. Spears) *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications,* pp. 81-96, L. Erlbaum Associates Inc., Broadway Hillsdale, NJ. doi: 10.5555/772072.772081

Brown, E., & Cairns, P. (2004). A grounded investigation of game immersion. In *CHI'04 extended abstracts on Human factors in computing systems*. ACM, New York, 1297-1300.

Candello, H., Pinhanez, C., & Figueiredo, F. (2017). Typefaces and the perception of humanness in natural language chatbots. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, New York, 3476-3487. doi: 10.1145/3025453.3025919

Candello, H., Pinhanez, C., Pichiliani, M., Cavalin, P., Figueiredo, F., Vasconcelos, M., & Do Carmo, H. (2019). The effect of audiences on the user experience with conversational interfaces in physical spaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ACM, New York, 1-13. doi: 10.1145/3290605.3300320

Cassell, J., Bickmore, T., Billinghurst, M., Campbell, L., Chang, K., Vilhjálmsson, H., & Hao, Y. (1999) Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, New York, 520–527. doi: 10.1.1.333.2333

Cervone, A., Gambi, E., Tortoreto, G., Stepanov, E. A., & Riccardi, G. (2018). Automatically predicting user ratings for conversational systems. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it* (Vol. 10, p. 12), Accademia University Press, Torino. doi: 10.4000/BOOKS.AACCADEMIA.3151

Chaix, B., Bibault, J. E., Pienkowski, A., Delamon, G., Guillemassé, A., Nectoux, P., & Brouard, B. (2019). When Chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a Chatbot. *Jmir Cancer*, *5*(1). doi: 10.2196/12856

Chattaraman, V., Kwon, W. S., Gilbert, J. E., & Ross, K. (2019). Should AI-Based, conversational digital assistants employ social-or task-oriented interaction style? A task-competency and reciprocity perspective for older adults. *Computers in Human Behavior*, *90*, 315-330. doi: 10.1016/j.chb.2018.08.048

Chung, M., Ko, E., Joung, H., & Kim, S. J. (2018). Chatbot e-service and customer satisfaction regarding luxury brands. *Journal of Business Research*, *117*, 587-595. doi: 10.1016/j.jbusres.2018.10.004

Ciechanowski, L., Przegalinska, A., & Wegner, K. (2017). The necessity of new paradigms in measuring human-chatbot interaction. In *Proceedings of the International Conference on Applied Human Factors and Ergonomics*. 205-214. Springer, Cham. doi: 10.1007/978-3-319-60747-4_19

Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., & Cowan, B. (2019a). The State of Speech in HCI: Trends, Themes and Challenges. In *Interacting with computers*, *31*(4), 349-371. Oxford University Press. doi: 10.1093/iwc/iwz016

Clark, L., Pantidi, N., Cooney, O., Doyle, D., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., Wade, V., & Cowan, B. R. (2019b). What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, Paper 475, 1–12. doi: 10.1145/3290605.3300705

Colby, K. M., Weber, S. & Hilf, F. D. (1971). Artificial Paranoia. *Artificial Intelligence*, *2*(1), 1-25. doi: 10.1016/0004-3702(71)90002-6

Corti, K., & Gillespie, A. (2016). Co-constructing intersubjectivity with artificial conversational agents: people are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior*, *58*, 431-442. doi: 431-442 10.1016/j.chb.2015.12.039

Cranshaw, J., Elwany, E., Newman, T., Kocielnik, R., Yu, B., Soni, S., Teevan, J. & Monroy-Hernández, A. (2017). Calendar. help: Designing a workflow-based scheduling agent with humans in the loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, New York, 2382-2393. doi: 10.1145/3025453.3025780

de Barcelos Silva, A., Gomes, M. M., da Costa, C. A., da Rosa Righi, R., Barbosa, J. L. V., Pessin, G., De Doncker, G. & Federizzi, G. (2020). Intelligent personal assistants: A systematic literature review. *Expert Systems with Applications*, *147*, 113193. doi: 10.1016/j.eswa.2020.113193

De Kleijn, R., Wijnen, M., & Poletiek, F. (2019). The effect of context-dependent information and sentence constructions on perceived humanness of an agent in a Turing test. *Knowledge-Based Systems, 163*, 794-799. doi: 10.1016/j.knosys.2018.10.006

de Medeiros, L. F., Junior, A. K., & Moser, A. (2019). A Cognitive Assistant that Uses Small Talk in Tutoring Conversation. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(11), 138-159. doi: 10.3991/ijet.v14i11.10288

Debeauvais, T. (2016). *Challenge and retention in games.* (Ph.D., University of California, Irvine). ProQuest Dissertations and Theses, Retrieved from http://search.proquest.com.libaccess.sjlibrary.org/docview/1808939056?accountid=10361. Accessed October 3, 2019.

DeHaemer, M. J., & Wallace, W. A. (1992). The effects on decision task performance of computer synthetic voice output. *International journal of man-machine studies*, *36*(1), 65-80. doi: 10.1016/0020-7373(92)90052-M

Denecke, K., Hochreutener, S. L., Pöpel, A. & May, R. (2018). Self- Anamnesis with a Conversational User Interface: Concept and Usability Study. *Methods of information in medicine*, *57*(05/06), 243-252. doi: 10.1055/s-0038-1675822

Denecke, K., Vaaheesan, S., & Arulnathan, A. (2020). A Mental Health Chatbot for Regulating Emotions (SERMO)-Concept and Usability Test. *IEEE Transactions on Emerging Topics in Computin*g. doi: 10.1109/TETC.2020.2974478

Derks, D., Fischer, A. H., & Bos, A. E. R. (2008). The role of emotion in computer mediated communication: A review. *Computers in Human Behavior*, *24*(3), 766–785. doi: 10.1016/j.chb.2007.04.004

Desmet, P., & Hekkert, P. (2007). Framework of product experience. *International journal of design*, *1*(1), 57-66.

Distler, V., Lallemand C. & Bellen, T. (2018) Acceptability and Acceptance of Autonomous Mobility on Demand: The Impact of an Immersive Experience, In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* Article 612, 1-10. ACM, New York, 1–10. doi: 10.1145/3173574.3174186

Doherty, K. & Doherty, G. (2018). Engagement in HCI: Conception, Theory and Measurement. *ACM Comput. Surv. 51*(5), Article 99 (November 2018), 39 pages. doi: 10.1145/3234149

Doyle, P. R., Edwards, J., Dumbleton, O., Clark, L., and Cowan, B. R. (2019). Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '19).* ACM, New York, Article 5, 1–12. doi: 10.1145/3338286.3340116

Egger, F. N., Helander, M., Khalid, K. M. & Tham, N. (2001). Affective design of e-commerce user interfaces: How to maximise perceived trustworthiness. In *Proceedings of the international conference on affective human factors design*, Asean Academic Press, London, 317-324.

Fadhil, A., Schiavo, G., Wang, Y., & Yilma, B. A. (2018). The effect of emojis when interacting with conversational interface assisted health coaching system. In *Proceedings of the 12th EAI international conference on pervasive computing technologies for healthcare*, 378-383, ACM, New York. doi: 10.1145/3240925.3240965

Fessl, A., Rivera-Pelayo, V., Müller, L., Pammer, V., Lindstaedt, S. (2011). Motivation and User Acceptance of Using Physiological Data to Support Individual Reflection. In *Proceedings of the 2nd and 3rd International Workshops on Motivational and Affective Aspects part of the European Conference on Technology Enhanced Learning (ECTEL)* CEUR WS Proceedings, Volume 957.

Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T., & Sprecher, S. (2012). Online dating: A critical analysis from the perspective of psychological science. *Psychological Science in the Public interest*, *13*(1), 3-66. doi: 10.1177/1529100612436522

Fiore, D., Baldauf, M., & Thiel, C. (2019). " Forgot your password again?" acceptance and user experience of a chatbot for in-company IT support. In *Proceedings of the 18th International Conference on Mobile and Ubiquitous Multimedia*, 1-11, ACM, New York. doi: 10.1145/3365610.3365617

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). 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), e19. doi: 10.2196/mental.7785

Følstad A., Skjuve M. & Brandtzaeg P.B. (2019). Different Chatbots for Different Purposes: Towards a Typology of Chatbots to Understand Interaction Design. *INSCI 2018. Lecture Notes in Computer Science*, vol 11551, 145-156. Springer, Cham. doi: 10.1007/978-3-030-17705-8_13

Følstad, A., & Skjuve, M. (2019). Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st International Conference on Conversational User Interfaces*, 1-9, ACM, New York. doi: 10.1145/3342775.3342784

Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? An exploratory interview study. In *Proceedings of the International Conference on Internet Science*, 194-208. Springer, Cham. doi: 10.1007/978-3-030-01437-7_16

Forrester Research. (2016). Chatbots 101: Building Conversational Interfaces. Technical report. Accessed October 15, 2020.

Frison, A. K., Wintersberger, P., Riener, A., Schartmüller, C., Boyle, L. N., Miller, E., & Weigl, K. (2019, May). In UX We Trust: Investigation of Aesthetics and Usability of Driver-Vehicle Interfaces and Their Impact on the Perception of Automated Driving. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13), ACM, New York. doi: 10.1145/3290605.3300374

Galko, L., Porubän, J., & Senko, J. (2018). Improving the User Experience of Electronic University Enrollment. In *Proceedings of the 16th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, 179-184, IEEE. doi: 10.1109/ICETA.2018.8572054

Gartner (2017). Making Live Chat a Must-Have Engagement Channel. Technical Report. Retrieved from https://www.gartner.com/en/documents/3803508/making-live-chat-a-must-have-engagement-channel. Accessed October 15, 2020.

Gasson, S. (2009). Employing a grounded theory approach for MIS research. In *Handbook of Research on Contemporary Theoretical Models in Information Systems* (Dwivedi, Y. K., Lal B., Williams, M. D., Schnebergerand, S. L., Wade, M., Eds), 34–56. IGI Global, Hershey, PA.

Gilbert, R. L., Forney, A. (2015). Can avatars pass the Turing test? Intelligent agent perception in a 3D virtual environment. *International Journal of Human-Computer Studies*, *73*, 30-36. doi: 10.1016/j.ijhcs.2014.08.001

Gilmartin, E., Collery, M., Su, K., Huang, Y., Elias, C., Cowan, B. R. & Campbell, N. (2017). Social talk: making conversation with people and machine. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents (ISIAA 2017).* ACM, New York, 31–32. doi: https://doi.org/10.1145/3139491.3139494

Glaser, B. G. & Strauss, A. L. (2017) *Discovery of Grounded Theory: Strategies for Qualitative Research*. Routledge, Evanston, IL.

Go, E., & Sundar, S. S. (2019). Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, *97*, 304-316. doi: 10.1016/j.chb.2019.01.020

Goethe O., Salehzadeh Niksirat K., Hirskyj-Douglas I., Sun H., Law E.LC., Ren X. (2019). From UX to Engagement: Connecting Theory and Practice, Addressing Ethics and Diversity. In *Proceedings of the International Conference on Human-Computer Interaction*, 91-99. Springer, Cham. doi: 10.1007/978-3-030-23560-4_7

Greer, S., Ramo, D., Chang, Y. J., Fu, M., Moskowitz, J., & Haritatos, J. (2019). Use of the chatbot "Vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: Randomized controlled feasibility trial. *JMIR mHealth and uHealth, 7*(10), e15018. doi: 10.2196/15018

Grudin, J., & Jacques, R. (2019). Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-11), ACM, New York. doi: 10.1145/3290605.3300439

Hassenzahl, M. (2003) The thing and I: understanding the relationship between user and product. In: Blythe M.A., Overbeeke K., Monk A.F., Wright P.C. (eds) Funology. Human-Computer Interaction Series, vol 3, 31-42. Dordrecht: Springer. doi: 10.1007/1-4020-2967-5_4

Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & information technology*, *25*(2), 91-97. doi: 10.1080/01449290500330331

Heider, F. (1946). Attitudes and Cognitive Organization. *The Journal of Psychology*, *21*(1), 107–112. doi: 10.1080/00223980.1946.9917275

Hekler, E. B., Klasnja, P., Froehlich, J. E., & Buman, M. P. (2013). Mind the theoretical gap: interpreting, using, and developing behavioral theory in HCI research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, 3307-3316. doi: 10.1145/2470654.2466452

Hentschel, E. (1999). Communication on IRC. *statistics*, *100*(100), 100, doi: 10.13092/lo.1.1084

Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in human behavior*, *49*, 245-250. doi: 10.1016/j.chb.2015.02.026

Ho, C. C., & MacDorman, K. F. (2017). Measuring the uncanny valley effect. *International Journal of Social Robotics*, *9*(1), 129-139. doi: 10.1007/s12369-016-0380-9

Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & McTear, M. (2019). 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, New York. doi: 10.1145/3335082.3335094

Huang, Y. K., Hsieh, C. H., Li, W., Chang, C., & Fan, W. S. (2019, December). Preliminary Study of Factors Affecting the Spread and Resistance of Consumers' Use of AI Customer Service. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*, 132-138, ACM, New York. doi: 10.1145/3375959.3375968

Hussain, S., Sianaki, O. A., & Ababneh, N. (2019). A survey on conversational agents/chatbots classification and design techniques. In *Workshops of the International Conference on Advanced Information Networking and Applications*, 946-956. Springer, Cham. doi: 10.1007/978-3-030-15035-8_93

ISO (1998). *ISO 9241: Ergonomic requirements for office work with visual display terminals (VDTs) Part 11: Guidance on usability*. Geneva: International Organization for Standardization.

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018). Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*, 895-906, ACM, New York. doi: 10.1145/3196709.3196735

Jin, Y., Cai, W., Chen, L., Htun, N. N., & Verbert, K. (2019). MusicBot: Evaluating critiquing-based music recommenders with conversational interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (pp. 951-960), ACM, New York. doi: 10.1145/3357384.3357923

Joinson, A. N. (2001). Self-disclosure in computer mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, *31*, 177–192. doi: 10.1002/ejsp.36

Jones, S. E., Miller, J. D., & Lynam, D. R. (2011). Personality, antisocial behavior, and aggression: A meta-analytic review. *Journal of Criminal Justice*, *39*(4), 329–337. doi: 10.1016/j.jcrimjus.2011.03.004.

Kashdan, T. B., Goodman, F. R., Mallard, T. T., DeWall, C. N. (2016). What Triggers Anger in Everyday Life? Links to the Intensity, Control, and Regulation of These Emotions, and Personality Traits. *Journal of Personality*, *84*(6), 737-749. doi: 10.1111/jopy.12214

Kennedy, C. M., Powell, J., Payne, T. H., Ainsworth, J., Boyd, A., & Buchan, I. (2012). Active assistance technology for health-related behavior change: an interdisciplinary review. *Journal of medical Internet research*, *14*(3), e80. doi: 10.2196/jmir.1893

Kim, S., Eun, J., Oh, C., Suh, B., & Lee, J. (2020). Bot in the Bunch: Facilitating Group Chat Discussion by Improving Efficiency and Participation with a Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1-13), ACM, New York. doi: 10.1145/3313831.3376785

Kimani, E., Rowan, K., McDuff, D., Czerwinski, M., & Mark, G. (2019). A Conversational Agent in Support of Productivity and Wellbeing at Work. In *Proceedings of the 8th International Conference on Affective Computing and Intelligent Interaction (ACII),* 1-7, IEEE. doi: 10.1109/ACII.2019.8925488

Klein, R. (2007). Internet-based patient-physician electronic communication applications: Patient acceptance and trust. *e-service Journal*, *5*(2), 27–51. doi: 10.2979/esj.2007.5.2.27

Klopfenstein, L. C., Delpriori, S., Malatini, S., & Bogliolo, A. (2017). The rise of bots: A survey of conversational interfaces, patterns, and paradigms. In *Proceedings of the 2017 conference on designing interactive systems*, 555-565, ACM, New York. doi: 10.1145/3064663.3064672

Kuramoto, I., Baba, J., Ogawa, K., Yoshikawa, Y., Kawabata, T., & Ishiguro, H. (2018). Conversational Agents to Suppress Customer Anger in Text-based Customer-support Conversations.

In *Proceedings of the 6th International Conference on Human-Agent Interaction* (pp. 114-121), ACM, New York. doi: 10.1145/3284432.3284457

Lallemand, C., Gronier, G. & Koenig, V. (2015). User experience: A concept without consensus? Exploring practitioners' perspectives through an international survey. *Computers in Human Behavior 43*, 35–48. doi: 10.1016/j.chb.2014.10.048

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, D., Magrabi, F., Lau, A. Y. S. & Coiera, E. (2018). Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, *25*(9), 1248-1258. doi: 10.1093/jamia/ocy072

Law, E. L. C., Vermeeren, A. P., Hassenzahl, M., & Blythe, M. (2007). Towards a UX manifesto. In *Proceedings of HCI 2007 The 21st British HCI Group Annual Conference*, 205–206, BCS Learning & Development, Swindon UK. doi: 10.5555/1531407.1531468

Leahu, L., Cohn, M. & March, W. (2013). How categories come to matter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13),* 3331–3334. ACM, New York. doi: https://doi.org/10.1145/2470654.2466455

Lee, J. R. & Nass, C. I. (2010). Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In J-E. R. Lee, & C. I. Nass (Eds), *Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives,* 1-15. IGI Global, Hershey, PA. doi: 10.4018/978-1-61520-901-9.ch001

Lee, M., Ackermans, S., van As, N., Chang, H., Lucas, E., & IJsselsteijn, W. (2019). Caring for Vincent: A Chatbot for Self-Compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing System*s, 1-13. ACM, New York. doi: 10.1145/3290605.3300932

Lee, S., Lee, N., & Sah, Y. J. (2020a). Perceiving a Mind in a Chatbot: Effect of Mind Perception and Social Cues on Co-presence, Closeness, and Intention to Use. *International Journal of Human–Computer Interaction*, *36*(10), 930-940. doi: 10.1080/10447318.2019.1699748

Lee, Y. C., Yamashita, N., Huang, Y., & Fu, W. (2020b). " 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*, 1-12, ACM, New York. doi: 10.1145/3313831.3376175

Li, C. H., Yeh, S. F., Chang, T. J., Tsai, M. H., Chen, K., & Chang, Y. J. (2020). A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-12, ACM New York. doi: 10.1145/3313831.3376209

Li, Y. & Yeh, Y. (2010) Increasing trust in mobile commerce through design aesthetics. *Computers in Human Behavior,* 26 (4), 673–684. doi: 10.1016/j.chb.2010.01.004

Liao, Q. V., Mas-ud Hussain, M., Chandar, P., Davis, M., Khazaeni, Y., Crasso, M. P., Wang, D., Muller, M., Shami, N. S. & Geyer, W. (2018). All Work and No Play?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1-13, ACM, New York. doi: 10.1145/3173574.3173577

Lindgaard, G., Dudek, C., Sen, D., Sumegi, L. & Noonan, P. (2011). An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction (TOCHI) 18* (1), 1-30. doi: 10.1145/1959022.1959023

Littell, J. H., Corcoran, J., & Pillai, V. (2008). *Systematic reviews and meta-analysis*. Oxford University Press, Oxford, UK.

Liu, D., Santhanam, R. & Webster, J. (2017). Toward Meaningful Engagement: A Framework for Design and Research of Gamified Information Systems. *MIS Quarterly, 41*(4), 1011-1039. doi: 10.25300/MISQ/2017/41.4.01

Liu, R., & Dong, Z. (2019). A study of user experience in knowledge-based QA chatbot design. In *Proceedings of International Conference on Intelligent Human Systems Integration*, 589-593. Springer, Cham. doi: 10.1007/978-3-030-11051-2_89

Liu, B., & Sundar, S. S. (2018). Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking*, *21*(10), 625-636. doi: 10.1089/cyber.2018.0110.

Luger, E., and Sellen, A. (2016). "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*, 5286–5297. ACM, New York. doi: 10.1145/2858036.2858288

Lukoff, K., Yu, C., Kientz, J. & Hiniker, A. (2018). What Makes Smartphone Use Meaningful or Meaningless? In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies,* 2(1), 1-26. doi: 10.1145/3191754.

Lunsford, R., & Oviatt, S. (2006). Human perception of intended addressee during computer-assisted meetings. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pp. 20-27 ACM, New York. doi: 10.1145/1180995.1181002

Ly, K. H., Ly, A. M., & Andersson, G. (2017). A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods. *Internet interventions*, *10*, 39-46. doi: 10.1016/j.invent.2017.10.002

Mahlke, S. (2008). *User Experience of interaction with technical systems. Theories, methods, empirical results, and their application to the design of interactive systems*. VDM Verlag, Saarbrücken, Germany.

Mahmood, T., & Ricci, F. (2009). Improving recommender systems with adaptive conversational strategies. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, 73-82. ACM, New York. doi: 10.1145/1557914.1557930

Mattarelli, E., Bertolotti, F., & Macrì, D. M. (2013). The use of ethnography and grounded theory in the development of a management information system. *European Journal of Information Systems*, *22*(1), 26-44. doi: 10.1057/ejis.2011.34

Mayer, R. C., Davis, J. H., Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709-734. doi: 10.5465/amr.1995.9508080335

Medeiros, L., Gerritsen, C., & Bosse, T. (2019). Towards humanlike chatbots helping users cope with stressful situations. In *Proceedings of the International Conference on Computational Collective Intelligence*, 232-243. Springer, Cham. doi: 10.1007/978-3-030-28377-3_19

Mehrabian, A. & Epstein N. (1972) A measure of emotional empathy. *Journal of Personality*, 40(4), 525–543. doi: 10.1111/j.1467-6494.1972.tb00078.x.

Mencarini, E., Rapp, A., Tirabeni, L., Zanacanaro, M. (2019). Designing Wearable Systems for Sport: A Review of Trends and Opportunities in Human-Computer Interaction. *IEEE Transactions on Human-Machine Systems*, *49*(4), 314-325. doi: 10.1109/THMS.2019.2919702

Mendez, S. L., & Conley, V. M., & Johanson, K., & Gosha, K., & Mack, N. A., & Haynes, C. L., & Gerhardt, R. A. (2019). The Use of Chatbots in Future Faculty Mentoring: A Case of the Engineering Professoriate. Paper presented at 2019 ASEE Annual Conference & Exposition, Tampa, Florida. doi: 10.18260/1-2--33434

Michaud, L. N. (2018). Observations of a new chatbot: drawing conclusions from early interactions with users. *IT Professional, 20*(5), 40-47. doi: 10.1109/MITP.2018.053891336

Mori, E., Takeuchi, Y., & Tsuchikura, E. (2019). How Do Humans Identify Human-Likeness from Online Text-Based Q&A Communication?. In *Proceedings of the International Conference on Human-Computer Interaction,* 330-339. Springer, Cham. doi: 10.1007/978-3-030-22646-6_24

Mori, M. (1970) Bukimi no tani - The uncanny valley (K. F. MacDorman & T. Minato,. Trans.) - *Energy*, 7(4), 33–35.

Morrissey, K., & Kirakowski, J. (2013). 'Realness' in chatbots: establishing quantifiable criteria. In *Proceedings of the International conference on human-computer interaction*, 87-96. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-39330-3_10

Mou, Y., & Xu, K. (2017). The media inequality: Comparing the initial human-human and human-AI social interactions. *Computers in Human Behavior*, *72*, 432-440. doi: 10.1016/j.chb.2017.02.067

Nadarzynski, T., Miles, O., Cowie, A., & Ridge, D. (2019). Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. *Digital health*, *21*(5). doi: 10.1177/2055207619871808

Nunes, F., Verdezoto, N., Fitzpatrick, G., Kyng, M., Grönvall, E., & Storni, C. (2015). Self-care technologies in HCI: Trends, tensions, and opportunities. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *22*(6), 1-45. doi: 10.1145/2803173

O'Brien, H. & Toms, E. (2008) What is user engagement? A conceptual framework for defining user engagement with technology. *JASIST*, *59*(6), 938-955. doi: 10.1002/asi.20801

Ortony, A., Clore, G.L., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge University Press, Cambridge, MA.

Park, M., Aiken, M., & Salvador, L. (2018). How do humans interact with chatbots?: An analysis of transcripts. *International Journal of Management & Information Technology*, *14*, 3338-3350. doi: 10.24297/ijmit.v14i0.7921

Pease, C. R., & Lewis, G. J. (2015). Personality links to anger: Evidence for trait interaction and differentiation across expression style. *Personality and Individual Differences*, *74*, 159–164. doi: 10.1016/j.paid.2014.10.018.

Pedrotti, M., & Nistor, N. (2016). User Motivation and Technology Acceptance in Online Learning Environments. In *Proceedings of the European Conference on Technology Enhanced Learning*, 472-477. Springer, Cham. doi: 10.1007/978-3-319-45153-4_45

Pereira, J., & Díaz, Ó. (2019). Using health chatbots for behavior change: a mapping study. *Journal of medical systems*, *43*(5), 135. doi: 10.1007/s10916-019-1237-1

Pérez-Marín, D., & Pascual-Nieto, I. (2013). An exploratory study on how children interact with pedagogic conversational agents. *Behaviour & Information Technology*, *32*(9), 955-964. doi: 10.1080/0144929X.2012.687774

Peters, C., Castellano, G. & de Freitas, S. (2009). An exploration of user engagement in HCI. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots (AFFINE '09),* Article 9, 1–3. ACM, New York. doi: 10.1145/1655260.1655269

Portela, M., & Granell-Canut, C. (2017). A new friend in our smartphone? Observing Interactions with Chatbots in the Search of Emotional Engagement. In *Proceedings of the XVIII International Conference on Human Computer Interaction*, 1-7. doi: 10.1145/3123818.3123826

Procter, M., Heller, R., & Lin, F. (2018). Classifying Interaction Behaviors of Students and Conversational Agents Through Dialog Analysis. In *Proceedings of the International Conference on Intelligent Tutoring Systems*, 373-379. Springer, Cham. doi: 10.1007/978-3-319-91464-0_42

Rapp, A. (2019). Design fictions for behaviour change: exploring the long-term impacts of technology through the creation of fictional future prototypes. *Behaviour & Information Technology, 38*(3), 244-272. doi: 10.1080/0144929X.2018.1526970.

Rapp. A. (2020). Design fictions for learning: A method for supporting students in reflecting on technology in Human-Computer Interaction courses. *Computers & Education*, *145*, 1-18. doi: 10.1016/j.compedu.2019.103725.

Ren, X. (2016) Rethinking the Relationship between Humans and Computers. *IEEE Computer, 49*(8), 104–108. doi: 10.1109/MC.2016.253

Rödel, C., Stadler, S., Meschtscherjakov, A. & Tscheligi, M. (2014). Towards autonomous cars: The effect of autonomy levels on acceptance and user experience. In *Proceedings of the 6th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '14),* 1–8. ACM, New York*.* doi: 10.1145/2667317.2667330

Ruan, S., Jiang, L., Xu, J., Tham, B. J. K., Qiu, Z., Zhu, Y., Burnane, E., Brunskill, E. & Landay, J. A. (2019). Quizbot: A dialogue-based adaptive learning system for factual knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, , 1-13. ACM, New York. doi: 10.1145/3290605.3300587

Saket, B., Endert, A. & Stasko, J. (2016) Beyond Usability and Performance: A Review of User Experience-focused Evaluations in Visualization. In *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization - BELIV '16*, 133–142. ACM, New York. doi: 10.1145/2993901.2993903

Sanny, L., Susastra, A., Roberts, C., & Yusramdaleni, R. (2020). The analysis of customer satisfaction factors which influence chatbot acceptance in Indonesia. *Management Science Letters*, *10*(6), 1225-1232. 10.5267/j.msl.2019.11.036

Schuetzler, R. M., Grimes, G. M., & Giboney, J. S. (2019). The effect of conversational agent skill on user behavior during deception. *Computers in Human Behavior*, *97*, 250-259. doi: 10.1016/j.chb.2019.03.033

Schuitema, G., Steg, L., & Forward, S. (2010). Explaining differences in acceptability before and acceptance after the implementation of a congestion charge in Stockholm. *Transportation Research Part A: Policy and Practice*, *44*(2), 99–109. doi: 10.1016/j.tra.2009.11.005

Schuler, D., & Namioka, A. (Eds.). (1993). *Participatory design: Principles and practices*. CRC Press, Boca Raton, FL.

Schumaker, R. P., & Chen, H. (2010). Interaction analysis of the alice chatterbot: A two-study investigation of dialog and domain questioning. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *40*(1), 40-51. doi: 10.1109/TSMCA.2009.2029603

Schwind, V., Wolf, K., & Henze, N. (2018). Avoiding the uncanny valley in virtual character design. *interactions*, *25*(5), 45-49. doi: 10.1145/3236673

Seering, J., Luria, M., Ye, C., Kaufman, G., & Hammer, J. (2020). It Takes a Village: Integrating an Adaptive Chatbot into an Online Gaming Community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-13. ACM, New York. doi: 10.1145/3313831.3376708

Sensuse, D. I., Dhevanty, V., Rahmanasari, E., Permatasari, D., Putra, B. E., Lusa, J. S., Misbah.M. & Prima, P. (2019). Chatbot Evaluation as Knowledge Application: a Case Study of PT ABC. In *2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE)* (pp. 1-6), IEEE. doi: 10.1109/ICITEED.2019.8929967

Shi, W., Wang, X., Oh, Y. J., Zhang, J., Sahay, S., & Yu, Z. (2020). Effects of Persuasive Dialogues: Testing Bot Identities and Inquiry Strategies. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1-13. ACM, New York. doi: 10.1145/3313831.3376843

Skjuve, M., Haugstveit, I. M., Følstad, A., & Brandtzaeg, P. B. (2019). Help! Is my chatbot falling into the uncanny valley? An empirical study of user experience in human-chatbot interaction. *Human Technology*, *15*(1), 30-54. doi: 10.17011/HT/URN.201902201607

Smestad, T. L. & Volden, F. (2018). Chatbot Personalities Matters: Improving the user experience of chatbot interfaces. In: Bodrunova S. et al. (Eds.) *Internet Science. INSCI 2018. Lecture Notes in Computer Science*, vol 11551, 170-181. Springer, Cham. doi: 10.1007/978-3-030-17705-8_15

Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the Facebook Messenger. *Computers & Education*, *151*. doi: 10.1016/j.compedu.2020.103862

Spears, R., & Lea, M. (1992). Social influence and the influence of the "social" in computer-mediated communication. In M. Lea (Ed.), *Contexts of computer-mediated communication* (pp. 30– 65). Wheatsheaf, Hemel Hempsted, UK.

Stein, J.P. & Ohler P. (2017). Venturing into the uncanny valley of mind—the influence of mind attribution on the acceptance of human-like characters in a virtual reality setting. *Cognition*, *160*, 43–50. doi: 10.1016/j.cognition.2016.12.010

Sullivan, L. (2017). Facebook chatbots hit 70% failure rate as consumers warm up to the tech. Retrieved from https://www.mediapost.com/publications/article/295718/facebookchatbots- hit-70-failure-rate-as-consumer.html. Accessed 21 February, 2020.

Sundar, S. S., Bellur, S., Oh, J., Jia, H., & Kim, H. S. (2016). Theoretical importance of contingency in human-computer interaction: effects of message interactivity on user engagement. *Communication Research*, *43*(5), 595-625. doi: 10.1177/0093650214534962

Svenningsson, N., & Faraon, M. (2019). Artificial Intelligence in Conversational Agents: A Study of Factors Related to Perceived Humanness in Chatbots. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*, 151-161. ACM, New York. doi: 10.1145/3375959.3375973

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., Decero, E. & Loggarakis, A. (2020). User Experiences of Social Support From Companion Chatbots in Everyday Contexts: Thematic Analysis. *Journal of medical Internet research*, *22*(3). doi: 10.2196/16235

Tärning, B., & Silvervarg, A. (2019). "I Didn't Understand, I´ m Really Not Very Smart"—How Design of a Digital Tutee's Self-Efficacy Affects Conversation and Student Behavior in a Digital Math Game. *Education Sciences*, *9*(3), 197. doi: 10.3390/EDUCSCI9030197

ter Stal, S., Kramer, L. L., Tabak, M., op den Akker, H., & Hermens, H. (2020). Design Features of Embodied Conversational Agents in eHealth: A Literature Review. *International Journal of Human-Computer Studies*, *138*. doi: 10.1016/j.ijhcs.2020.102409

Tidwell, L. C., & Walther, J. B. (2002). Computer-mediated communication effects on disclosure, impressions, and interpersonal evaluations: Getting to know one another a bit at a time. *Human Communication Research*, *28*(3), 317–348. doi: 10.1111/j.1468-2958.2002.tb00811.x

Trivedi, J. (2019). Examining the customer experience of using banking Chatbots and its impact on brand love: The moderating role of perceived risk. *Journal of internet Commerce*, *18*(1), 91-111. doi: 10.1080/15332861.2019.1567188

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, 59*, 433-460.

Urakami, J., Moore, B. A., Sutthithatip, S., & Park, S. (2019). Users' Perception of Empathic Expressions by an Advanced Intelligent System. In *Proceedings of the 7th International Conference on Human-Agent Interaction*, 11-18, ACM, New York. doi: 10.1145/33495 37.3351895

Väätäjä, H., Koponen, T. & Roto, V. (2009). Developing practical tools for user experience evaluation: a case from mobile news journalism. In *European Conference on Cognitive Ergonomics: Designing beyond the Product---Understanding Activity and User Experience in Ubiquitous Environments* (p. 23). VTT Technical Research Centre of Finland. doi: 10.5555/1690508.1690539

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, *64*(7), 456-464. doi: 10.1177/0706743719828977

Valkenburg, P. M., & Peter, J. (2009). The effects of instant messaging on the quality of adolescents' existing friendships: A longitudinal study. *Journal of Communication, 59*, 79–97. doi: 10.1111/j.1460-2466.2008.01405.x

Van den Broeck, E., Zarouali, B., & Poels, K. (2019). Chatbot advertising effectiveness: When does the message get through?. *Computers in Human Behavior*, *98*, 150-157 10.1016/j.chb.2019.04.009

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human bot interactions: Detection, estimation, and characterization, In *Proceedings of the Eleventh international AAAI conference on web and social media*, 280-289, AAAI Press, Palo Alto, California.

Völkel, S. T., Haeuslschmid, R., Werner, A., Hussmann, H., and Butz, A. (2020). How to Trick AI: Users' Strategies for Protecting Themselves from Automatic Personality Assessment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. ACM, New York, 1–15. doi: 10.1145/3313831.3376877

Vossen, H. G., Piotrowski, J.T. & Valkenburg, P.M. (2015). Development of the adolescent measure of empathy and sympathy (AMES). *Personality and Individual Differences*, *74*, 66–71. doi: 10.1016/j.paid.2014.09.040

Wallace, R. S. (2009) The anatomy of ALICE. In *Parsing the Turing Test*. Springer, 181–210. doi: 10.1007/978-1-4020-6710-5_13

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication research*, *23*(1), 3-43. doi: 10.1177/009365096023001001

Wang, X., & Nakatsu, R. (2013). How do people talk with a virtual philosopher: Log Analysis of a real-world application. In *International Conference on Entertainment Computing*, 132-137. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-41106-9_16

Warwick, K., & Shah, H. (2015). Intelligent agents: Conversations from human-agent imitation games. In D. H. J. van, J. Filipe, S. Loiseau, & B. Duval (Eds.), *ICAART 2015 - 7th International Conference on Agents and Artificial Intelligence, Proceedings* (Vol. 2, pp. 261-268). SciTePress, Setúbal.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45. doi: 10.1145/365153.365168

Werry, C. C. (1996). *Linguistic and interactional features of internet relay chat*. John Benjamins Publishing, Philadelphia, PA.

Westerman, D., Cross, A. C., & Lindmark, P. G. (2019). I believe in a thing called bot: Perceptions of the humanness of "chatbots". *Communication Studies*, *70*(3), 295-312. doi: 10.1080/10510974.2018.1557233

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. (2013) Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems*, *22*(1). 45-55, doi: 10.1057/ejis.2011.51

Wright, P., McCarthy, J., & Meekison, L. (2003) Making sense of experience. In: Blythe M.A.,
Overbeeke K., Monk A.F., Wright P.C. (Eds.). Funology. Human-Computer Interaction Series, vol 3,
43-53. Springer, Dordrecht. https://doi.org/10.1007/1-4020-2967-5_5

Xiao, Z., Zhou, M. X., Chen, W., Yang, H., & Chi, C. (2020). If I Hear You Correctly: Building and
Evaluating Interview Chatbots with Active Listening Skills. In *Proceedings of the 2020 CHI
Conference on Human Factors in Computing Systems*, 1-14. ACM, New York. doi:
10.1145/3313831.3376131.

Xu, A., Liu, Z., Guo, Y., Sinha, V., & Akkiraju, R. (2017). A new chatbot for customer service on
social media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*,
3506-3510. ACM, New York. doi: 10.1145/3025453.3025496

Yan, M., Castro, P., Cheng, P. & Ishakian, V. (2016). Building a Chatbot with Serverless Computing.
In *Proceedings of the 1st International Workshop on Mashups of Things and APIs (MOTA '16),*
Article 5, 1–4. ACM, New York. doi: 10.1145/3007203.3007217

Yen, C., & Chiang, M. C. (2020). Trust me, if you can: a study on the factors that influence
consumers' purchase intention triggered by chatbots based on brain image evidence and self-reported
assessments. *Behaviour & Information Technology*, 1-18. doi: 10.1080/0144929X.2020.1743362

Zamora, J. (2017). I'm sorry, dave, I'm afraid I can't do that: Chatbot perception and expectations. In
Proceedings of the 5th International Conference on Human Agent Interaction, 253-260. doi:
10.1145/3125739.3125766

Zarouali, B., Van den Broeck, E., Walrave, M., & Poels, K. (2018). Predicting consumer responses to
a chatbot on Facebook. *Cyberpsychology, Behavior, and Social Networking, 21*(8), 491-497. doi:
10.1089/cyber.2017.0518

Zaroukian, E., Bakdash, J. Z., Preece, A., & Webberley, W. (2017). Automation bias with a
conversational interface: User confirmation of misparsed information. In *2017 IEEE Conference on*

Cognitive and Computational Aspects of Situation Management (CogSIMA), 1-3, IEEE. doi: 10.1109/COGSIMA.2017.7929605

Zhang, Z., Cheng, W. & Gu, Z. (2016). User Experience Studies Based on Expectation Dis-confirmation Theory. In Marcus A. (Eds.) *Design, User Experience, and Usability: Novel User Experiences. DUXU 2016. Lecture Notes in Computer Science*, vol 9747, 670-677. Springer, Cham. https://doi.org/10.1007/978-3-319-40355-7_65

Zhou, M. X., Mark, G., Li, J., & Yang, H. (2019). Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS), 9*(2-3), 1-36. doi: 10.1145/3232077

Zimmerman, J. & Forlizzi, J. (2008) The Role of Design Artifacts in Design Theory Construction. *Artifact: Journal of Design Practice*, *2*(1), 41-45. doi: 10.1080/17493460802276893

Zumstein, D., & Hundertmark, S. (2017). Communicating and transacting with chatbots: insights from public transport. In *Proceedings of the 16th International Conference Applied Computing WWW/Internet 2017,* 55-62. IADIS.