UX Research on Conversational Human-AI Interaction: A Literature Review of the ACM Digital Library

Qingxiao Zheng
School of Information Sciences,
University of Illinois at Urbana-Champaign
USA
qzheng14@illinois.edu

Yiliu Tang
School of Informatics,
University of Illinois at Urbana-Champaign
USA
yiliut2@illinois.edu

Yiren Liu
School of Informatics,
University of Illinois at Urbana-Champaign
USA
yirenl2@illinois.edu

Weizi Liu
College of Media,
University of Illinois at Urbana-Champaign
USA
weizil2@illinois.edu

Yun Huang
School of Information Sciences,
University of Illinois at Urbana-Champaign
USA
yunhuang@illinois.edu

ABSTRACT

Early conversational agents (CAs) focused on *dyadic* human-AI interaction between humans and the CAs, followed by the increasing popularity of *polyadic* human-AI interaction, in which CAs are designed to mediate human-human interactions. CAs for *polyadic* interactions are unique because they encompass hybrid social interactions, i.e., human-CA, human-to-human, and human-to-group behaviors. However, research on *polyadic* CAs is scattered across different fields, making it challenging to identify, compare, and accumulate existing knowledge. To promote the future design of CA systems, we conducted a literature review of ACM publications and identified a set of works that conducted UX (user experience) research. We qualitatively synthesized the effects of *polyadic* CAs into four aspects of human-human interactions, i.e., communication, engagement, connection, and relationship maintenance. Through a mixed-method analysis of the selected *polyadic* and *dyadic* CA studies, we developed a suite of evaluation measurements on the effects. Our findings show that designing with social boundaries, such as privacy, disclosure, and identification, is crucial for ethical *polyadic* CAs. Future research should also advance usability testing methods and trust-building guidelines for conversational AI.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **HCI design and evaluation methods**; **User studies**.

KEYWORDS

Conversational Agent, Chatbot, Conversational AI, UX Research, Literature Review

ACM Reference Format:
Qingxiao Zheng, Yiliu Tang, Yiren Liu, Weizi Liu, and Yun Huang. 2022. UX Research on Conversational Human-AI Interaction: A Literature Review of the ACM Digital Library. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3491102.3501855

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501855

1 INTRODUCTION

There is a rapidly growing body of literature on conversational agents or chatbots [4]. As promising Artificial intelligence (AI) technologies, conversational agents are defined as "software that accepts natural language as input and generates natural language as output, engaging in a conversation with the user" [57]; chatbots, meanwhile, are computer programs designed to simulate conversation with human users via text [3, 4]. As these two terms are often perceived as interchangeable [105, 139], in the remainder of this paper, we refer to both conversational agents and chatbots as CAs. Scholars have shown that these machines are able to compensate for human shortcomings or exceed human capacities [55, 63, 180]. However, prior works focus on designing and evaluating *dyadic* human-AI interaction, which involve only one-to-one interactions between humans and their CAs [7, 15, 83, 150, 185]; whereas more recent works start tapping into *polyadic* human-AI interactions that also support human-human interactions [12, 76, 77, 168, 179].

Even though there are extensive literature reviews on CAs, e.g., [39, 44, 88, 113, 153], they do not address how *polyadic* CAs are designed and evaluated, nor present the effects of using *polyadic* CAs on handling the challenges of human-human interaction [67]. In this paper, we overview UX (user experience) research on *polyadic* CAs that 1) interact with more than one user in the same conversation and 2) engage in bidirectional conversations between all parties (human-AI and human-human). These *polyadic* CAs encompass a wide variety of complexities that *dyadic* Human-AI may not encounter, including multi-party interactions, social roles taking,

group hierarchy, or social tension [169]. To evaluate the effects of CAs' support in human-human interactions, we need to examine both human-CA behaviors and human-to-human behaviors, as well as human-to-group behaviors potentially. Given the unique challenges of the design space, little is known about how *polyadic* CAs should be designed to address these different aspects of social interaction and how they can positively influence the overall user experience [67, 153].

To fill the void, we conducted a literature review using a mixed-method approach, mapping out what exists in *polyadic* CAs research. Specifically, we screened 1,302 ACM papers and identified 36 papers that designed *polyadic* CAs and 135 that designed *dyadic* CAs, which included user evaluation results. We investigated what fundamental human-human interaction challenges are addressed by these works, what effects on human-human interaction are evaluated, and what issues are overlooked when designing these CAs.

The contributions of this work to HCI and social computing are manifold. First, we summarize the fundamental challenges (i.e., consensus reaching, uneven participation, lack of emotional awareness, etc.) of human-human interaction that *polyadic* CAs are designed to tackle and their promising results. Second, we synthesize the current practices (e.g., design originality, relationship types, social scale, evaluation method, etc.) and areas that fall short of empirical studies. Third, we point out the issues that are overlooked by the researchers and practitioners in these works and propose to design CAs with *boundary awareness* for supporting human-human interaction via conversational AI. Fourth, we present the differences in research interests and evaluation metrics between *dyadic* and *polyadic* CA studies and identify existing gaps and potential directions. Further, we envision several research opportunities on conversational human-AI interaction concerning the theoretical groundings, relational dynamics, functional dimensions, and metaphysical implications.

2 RELATED WORK

In this section, we first review briefly the history and the research landscape of CAs (Conversational Agents or Chatbots), and then point out the gaps in the existing literature.

2.1 Landscape of CAs

The current landscape of CAs is complex. The history of conversational interfaces could be traced back to the 1960s when text-based dialogue systems were designed to answer questions and for simulated casual conversation [106]. Since then, various terms have been used, and terminologies have become overlapping [139]. Historically, several distinctions have emerged in research around the terms, such as text-based and spoken dialogue systems, voice user interfaces, chatbots, embodied conversational agents, and social robots and situated agents [106]. In recent years, with the commercialization of CAs, more terms emerged. For example, the website chatbots.org references over 1,300 chatbots across nine categories [31] and provides 161 conversational AI synonyms [105].

Similarly to the terminologies, the typologies of CAs are also not univocally categorized [139]. In the past decade, researchers classified CAs based on multiple criteria, such as interaction modality (text-based, voice-based, mixed) [20, 69, 88, 116]; scale of social engagement (one-to-one, broadcasting, and community-based) [153]; knowledge domain (open domain or closed domain) [20, 69]; goals (task-oriented or non-task oriented) [56, 69]; duration and locus of control [53]; embodiment [32]; design approach (rule-based, retrieval based, and generative based) [69]; platform (mobile, laptop, web browser, SMS, cells, multimodal platforms) [79, 88], and application domains [44, 143].

In this paper, we include the aforementioned terminologies as our study objects, given that the CAs' major functionalities are conversational-based. For example, embodied conversational agents (ECAs) are agents simulating humans' face-to-face conversations and can use their body language when talking to users [32]. Four properties distinguish ECAs from CAs: ECAs can recognize and respond to both verbal and nonverbal user inputs; deliver both verbal and nonverbal outputs; deal with conversational functions; and give signals that help the conversations [32]. This work includes the ECA papers that conduct UX research on conversational interaction.

2.2 Gaps in Existing Literature Reviews of CAs

Most of the existing empirical studies focus on designing *dyadic* CAs [2, 42, 91, 148, 158], without addressing the unique design challenges of *polyadic* CAs. Thus, existing literature reviews address CAs' properties mainly in the context of *dyadic* conversationalists and the effect of interaction between the Human and CAs [153]. They look at how CAs can increase user engagement, enhance user experience, or enrich the relationship between humans and CAs. For example, one work mapped relevant themes in text-based CAs to understand user experiences, perceptions, and acceptances towards CAs [139]; and one categorized the characteristics of human-CA interactions into conversational, social, and personification characteristics [38]. Others are more specific, such as studying CAs' emotional intelligence [127], personalization [80], trust-building [142], and human-likeness [170].

Recently, an increasing number of CAs are designed to support *polyadic* human-AI interaction, where human-human communication is supported by CAs. While *dyadic* CAs are commonly used to communicate with individuals as personal assistants [112, 151], customer service agents [100, 157], and personal healthcare partners [16, 17, 58], some emotional intelligent CAs are designed to help team members gain awareness of other group members' emotional changes, report the overall sentiment of each group discussion, and maintain positive emotions during collaborations [12, 132]. Such *polyadic* CAs address unique social challenges that the *dyadic* CAs were not designed to meet, e.g., in the context of a group or a community. However, little is known about the empirical use and the effect of *polyadic* CAs on different scenarios of human-human interaction [40, 66, 67, 153].

Also, researchers and practitioners have increasingly noted a lack of understanding in achieving quality UX for CAs. Although the popularity of CA applications exists in multiple domains, studies repeatedly reported CAs causing both pragmatic issues, in which chatbots failed to understand or to help users achieve their intended goals [50, 51, 93], and issues in which CAs failed to engage users

over time [51, 93]. Thus, we set the survey scope to include papers that conducted user evaluations.

Our review of the CAs designed for *polyadic* Human-AI interaction can help researchers and practitioners identify common practices, research gaps, lessons learned, and promising areas for future advancements. Specifically, we are interested in addressing the following research questions:

RQ1: *What fundamental challenges of human-human interaction are addressed by these CAs?*

RQ2: *What are the research interests in dyadic and polyadic CAs?*

RQ3: *What are the practices of designing the CAs for polyadic human-AI interaction?*

RQ4: *What are the effects of using the proposed CAs on human-human interactions?*

RQ5: *What evaluation metrics have been used in understanding user experience?*

RQ6: *What issues are overlooked by scholars when designing polyadic CAs?*

3 METHOD

We used a literature review method that has been widely applied by prior works in HCI, e.g., [108, 122, 139], which has four stages: 1) Define: proposing the inclusion and exclusion criteria and identifying the appropriate data sources; 2) Search: developing specific query and collecting the papers through the data source; 3) Select: checking the search results against the inclusion and exclusion criteria and identifying the final papers for both *dyadic* and *polyadic* works; and 4) Analyze: examining the selected papers by applying a mixed-method approach. We illustrate the four steps in Figure 1 and present the details of each stage below.

3.1 Definition

In this section, we present how we defined the selection criteria and how we identified the appropriate data source.

3.1.1 Inclusion and Exclusion. A series of criteria were developed by researchers through multiple rounds of discussions. Criteria were selected to include works that were the most representative of the scope of our study(i.e., UX research on conversational human-AI interaction) and to filter irrelevant works. We employed the following inclusion and exclusion criteria.

For *dyadic* papers, we defined the inclusion criteria as follows: 1) selected articles need to study CAs that interact with only one human user in a session; 2) the CA interaction in the article is bidirectional; 3) users of the CA are aware of the existence of the CA; 4) the articles are research papers with user studies; 5) the major design feature is conversation-based, e.g., excluded sensory-based CA [24]; 6) the articles are written in English; 7) the articles are included in the selected database. Articles that only assessed CA task performance without meaningfully exploring the interaction experience of Human-CA as a conversational technology were excluded.

For *polyadic* papers, the inclusion criteria shared 2)-7) requirements of selecting *dyadic* papers', except that the selected articles need to study CAs that interact with greater than one human user, instead of with only one human user. Exclusion criteria were defined as: 1) articles that only assessed CA task performance without meaningfully exploring the interaction experience of Human-CA or Human-Human as a conversational technology; and 2) the CAs in the articles interacted with multiple users, but there were no human-human interactions.

3.1.2 Source. To identify the data source, we first randomly retrieved 200 papers from five databases: ACM Digital Library, IEEE, Web of Science, Scopus, and Science direct. After exploring the initial search results, we chose Association for Computing Machinery Digital Library (ACM DL) as the final source for our literature review. Two co-authors reviewed 40 papers from each source, using the inclusion and exclusion criteria among all five databases. The qualification rates were ACM DL (23.5%), IEEE (12.5%), Web of Science (14.6%), Scopus (17%), and Science direct (4.9%). Specifically, 52% IEEE papers were technical papers without user evaluations; 59% Web of Science papers contained no CA designs. Due to the low qualification rates of other sources, we ultimately decided to choose ACM DL as the data source for this literature review. This decision was also made by considering that the ACM DL features a wide selection of reliable HCI works, and has been used as a solo source for literature review works, e.g., [138, 175].

3.2 Search

The search query is composed of two parts. The first part of the search covers synonyms of conversational agents, and the second part specifies terms that may be relevant for identifying the *polyadic* papers. The list of search terms was inspired by several CA research reviews [44, 88, 139, 165] and was developed through several iterations and refinement by the research team, in line with prior HCI work.

More specifically, the search query included 15 terms, which are "conversational agent", "conversational AI", "intelligent assistant", "intelligent agent", "chatbot", "chatterbot", "chatterbox", "socialbot", "digital assistant", "conversational UI", "conversational interface", "conversation system", "conversational system", "dialogue system", and "dialog system". To explore *polyadic* works from the retrieved results, we searched 13 terms, "human-human," "human human," "multi-user," "multi-users," "multi user," "multi users," "multi-party," "multi-parties," "multi party," "multiparty-based", "multi parties", "multi model", and "multi-model."

3.2.1 Data Preparation. We searched the papers through two steps, as illustrated in Figure 1. First, we searched on Mar 3^{rd}, 2021 by using the query and connectors throughout the ACM DL, which resulted in 1,302 CA papers after removing duplicates. Second, we web-crawled the metadata and PDFs of these papers and built our database. To improve the stability and efficiency of the crawling process, we also saved the paper DOIs into a file and then crawled the metadata and the PDFs of the respective DOIs in sequence.

3.2.2 Database. The database consisted of two parts. The first part was a CSV file that comprehensively covers the metadata of the papers. It included the ten columns: paper DOI, title, authors, abstract, publication date, source, publisher, citations, keywords, affiliation of authors. The second part included the PDFs of the papers, as well as their corresponding text files. We used the Fitz module within the PyMuPDF project [73] to transform the PDF to

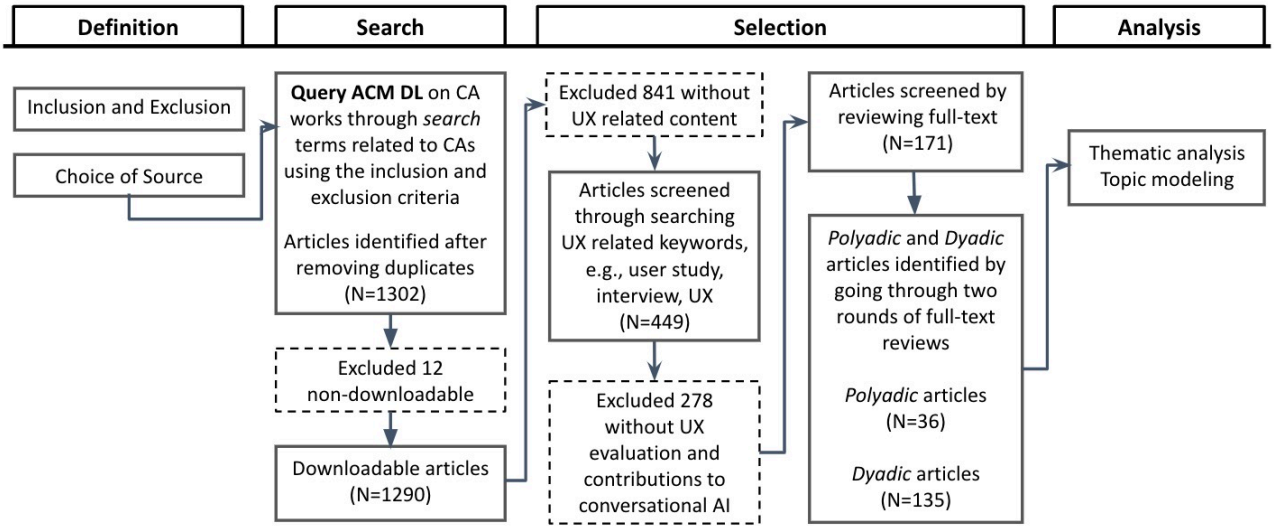

Figure 1: A Workflow Diagram of the Literature Review Process

text, as the tool has been applied in a number of works [36, 128, 129, 186].

3.3 Selection

Once having the database, we screened the paper records to check the *dyadic* and *polyadic* CA works that meet our criteria through two steps, as illustrated in Figure 1. We looked for papers that potentially had user studies by searching through the full texts of all papers with the keywords "user study", "user studies", "interview", "interviews" and "user experience", which led to 449 papers. To initiate the set of *polyadic* papers, we also searched through all the papers using the 13 key terms, e.g., "human-human," "human human," "multi-user," which resulted in an initial set of *polyadic* with 27 papers. We then removed duplicates between the two screening results. There were six (1.3%) duplicates between the two stages, resulting in 470 unique full texts. We then further reviewed the full texts against the identified inclusion and exclusion criteria for *polyadic* and *dyadic* papers. If ambiguities persisted about the eligibility of a specific paper, input was sought from the third co-author. The final study collected 36 (21.1%) *polyadic* papers and 135 (78.9%) *dyadic* papers.

3.4 Analysis

To examine the research landscape of UX research on conversational human-AI interaction and the differences between *polyadic* and *dyadic* works published in ACM DL, we analyzed the selected papers by applying a mixed-method approach.

3.4.1 Thematic Analysis. Prior literature review works [20, 69, 88, 139] have proposed several aspects when examining conversational AI research, e.g., regarding interaction modality [69], characteristics of embodiment [20] , application domains [139], evaluation methods [139], and social scale [20]. We found that these could be leveraged to code the research practices of *polyadic* CAs. However, the prior schemes were not sufficient for us to categorize the unique aspects of human-human interactions supported by *polyadic* CAs, e.g., fundamental challenges addressed, proven effects, and remaining issues. Thus, the authors adopted the grounded theory approach [181] and coded the attributes of the works using thematic analysis [27]. Two co-authors started reviewing the papers and did open coding of all papers independently, and then discussed and compiled their codes together. These codes were added as new attributes to address the proposed RQs. After reaching an agreement regarding the codes, two co-authors coded the rest papers separately, compared their codes, and discussed possible revisions [104, 135]. A final inter-rater reliability of 91% was achieved and deemed satisfactory [64].

3.4.2 Topic Modeling. Besides coding the research challenges and issues that are overlooked by scholars in designing *polyadic* CAs, we also applied topic modeling to explore the differences in discussion topics between the two groups of papers. The topic modeling method has been widely applied in prior works [60, 61, 162]. This method can be employed as follows. We started text pre-processing by removing punctuation marks, symbols, and stop-words. We then tokenized and lemmatized the text for regularization. We used lemmatization to group words with similar semantic meanings but different syntactic forms (such as plurality and tense). We experimented with both lemmatization and no lemmatization, and the results with lemmatization exhibited higher interpretability and were more coherent when we sampled the papers to explain the results. We then used non-negative matrix factorization (NMF) to conduct topic modeling based on the TF-IDF scores. We built two topic models separately for *dyadic* and *polyadic* works to explore the major topics discussed in either set of the papers. This provided closer insights into the frequent topics discussed by *dyadic* and *polyadic* CA papers. All text analysis steps were completed using Python (NLTK and Scikit-learn libraries).

4 FINDINGS

In this section, we present the findings of the ACM papers, 36 *polyadic* and 135 *dyadic* works, in order of the proposed RQs. Specifically, we show the fundamental challenges (RQ1 in Section 4.1), research interests (RQ2 in Section 4.2), practices (RQ3 in Section 4.3), proven effects (RQ4 in Section 4.4), evaluation metrics (RQ5 in Section 4.5), and issues (RQ6 in Section 4.6).

4.1 Addressing Fundamental Challenges of Human-Human Interaction (RQ1)

To answer our first research question, we identified major challenges in human-human interaction addressed in *polyadic* papers. We found that only some papers explicitly stated these challenges as pain points that motivated their designs. Most *challenges* addressed by *polyadic* CAs are in the collaborative learning, work, or group discussion contexts.

4.1.1 Inefficient Communication. As mentioned in prior research regarding group work and collaborations, several issues in human-human communication affect the efficiency. For example, in online group settings, team communications are often unstructured and less organized, with procrastination and distractions. Reaching consensus can be time-consuming for deliberative discussions [76, 77], and thoughts and discussions can be hard to organize and make sense of [189] and challenging to control in terms of conversational flow [21–23]. In terms of workflow, task management such as refining, assigning, and tracking can be difficult [168]; heavy workloads related to coordinating schedules among multiple parties can be tedious [42]; and switching between tools and platforms during collaborations can be burdensome [9]. Therefore, *polyadic* CAs are used to provide workflow support and discussion support to improve team communication efficiency.

4.1.2 Lack of Engagement. Inactive engagement is a common issue in multi-user interactions in collaborative teams. For example, in peer learning settings, it is challenging to engage students in provocative discussions [46], to promote productive talk among students, such as explaining to peers and re-voicing others' statements [48, 164], and to provide effective learning supports to others, e.g., by exhibiting an agreeable attitude and precipitating tension release [5, 37, 85, 86]. This is also true for collaborative work and online community contexts, in which more inputs [149] and greater community activity are needed [1, 101, 152, 154, 179]. Further issues of engagement in these contexts include uneven participation, in which a balanced amount of contribution is hard to achieve [76, 155], and user attractions are challenging to capture in casual social settings [124, 125, 188, 190].

4.1.3 Barriers in Relational Maintenance. One major challenge in multi-user social settings is maintaining positive relationships, which is crucial to forming a solid team or group. For example, in online collaborative work, there could be a lack of emotional awareness and mutual understanding between team members, as it is hard for them to detect and regulate emotions [12, 119, 132]. Moreover, it is always important to grow trust within a team, and the feasibility of using CAs for trust-building [161] and setting privacy boundaries [102] are explored and developed.

4.1.4 Need for Building Connections. In human-human interactions, there is a need for building social connections and identifying common grounds. This is particularly important in the "ice-breaking" and transitioning stages [99]. Also, people may seek similarities and agreement with their communication partners [70, 118] or need idea supports during chats [66]. This challenge could be more difficult in cross-cultural conversations. When people from different backgrounds meet for the first time and try to get acquainted with each other, it can be challenging to form impressions of each other free from the influence of cultural stereotypes [70].

4.2 Research Interests in *Dyadic* and *Polyadic* CAs (RQ2)

4.2.1 Research Interest over Time. The ACM papers we collected included UX research on conversational human-AI interaction starting from early 2000, e.g., *polyadic* CAs by Isbister et al. [70] and *dyadic* CAs by Chai et al. [34]. The works collected were not evenly distributed over the years, e.g., about four UX papers in 2010 and declining to one paper in 2012 for both. However, starting from 2018, there was a surge in the number of selected articles, with the number of *dyadic* papers reaching a peak of 48 in 2020 and the number of *polyadic* papers reaching a peak of nine in the same year.

4.2.2 Authors' Affiliations. There was a total of 596 authors for the 135 *dyadic* papers, averaging 4.4 authors per paper, and 379 (63.6%) out of the 596 authors were from academia. In comparison, there were a total of 145 authors for the 36 *polyadic* papers, averaging four authors per paper; and 121 (83.4%) out of the 145 authors were from academia.

4.2.3 Impactful Works in the Area. As shown in Table 1, our database collected a good body of *dyadic* papers. For example, Medhi et al. [107] evaluated the usability of a task-oriented CA for novice and low-literacy users and was highly cited, followed by several other usability studies of CAs [71, 72, 90] on improving health [15, 96], answering questions [95, 98], and counseling [150, 159] via CAs. Similarly, a sample of *polyadic* papers received high citations, as shown in Table 2. For example, Isbister et al. [70] designed a virtual agent to support human-human communication in virtual environments. Unlike the *dyadic* works, many selected *polyadic* works conducted UX research in the context of group environments, e.g., virtual meetings [70, 118], games [21, 23], online communities [149], and group collaboration [9, 189]. Our dataset included both *dyadic* and *polyadic* on improving work productivity [42, 78, 81, 155, 168, 173] and promoting education [5, 48, 85, 163]. The average number of citations of a *polyadic* paper in our database was 19 (SD=23) from ACM and 44 (SD=46) from Google Scholar. Five (13.9%) out of the 36 *polyadic* papers had no citations yet. The average number of citations of a *dyadic* paper in our database was 10 (SD=14) from ACM and 19 (SD=22) from Google Scholar, on December 21st, 2021. The dataset included works demonstrating impacts at different levels.

4.3 Practices of Designing CAs for *Polyadic* Human-AI Interaction (RQ3)

In the following section, we present aspects that are often shared by *polyadic* and *dyadic* CA works and aspects that are unique to *polyadic* CA works.

Table 1: A Sample of *Dyadic* ACM Papers that Conducted UX Research.

Year	Authors	ACM Citations	Google Scholar Citations
2011	Medhi, I., Patnaik, S., Brunskill, E., Gautama, S. N. N., Thies, W., & Toyama, K. [107]	166	275
2018	Jain, M., Kumar, P., Kota, R., & Patel, S. N. [72]	75	187
2013	Lisetti, C., Amini, R., Yasavur, U., & Rishe, N. [96]	85	168
2005	Bickmore, T. W., Caruso, L., & Clough-Gorr, K. [15]	62	143
2009	Schulman, D., & Bickmore, T. [150]	50	99
2018	Liao, Q. V., Hussain, M. M., Chandar, P., Davis, M., Khazaeni, Y., Crasso, M. P., Wang, D. K., Muller, M., Shami, N. S., & Geyer, W. [95]	37	48
2019	Kim, S., Lee, J., & Gweon, G. [78]	35	84
2017	Vtyurina, A., Savenkov, D., Agichtein, E., & Clarke, C. L. A. [173]	30	81
2019	Lovato, S. B., Piper, A. M., & Wartella, E. A. [98]	32	65
2018	Kocielnik, R., Avrahami, D., Marlow, J., Lu, D., & Hsieh, G. [81]	47	46
2010	Smyth, T. N., Etherton, J., & Best, M. L. [159]	37	86
2015	Tanaka, H., Sakti, S., Neubig, G., Toda, T., Negoro, H., Iwasaka, H. & Nakamura, S. [163]	31	48
2019	Zhou, M. X., Mark, G., Li, J., & Yang, H. [192]	30	54
2018	Jain, M., Kota, R., Kumar, P., & Patel, S. N. [71]	20	54
2010	Lau, T., Cerruti, J., Manzato, G., Bengualid, M., Bigham, J. P., & Nichols, J. [90]	25	44
2005	Marti, S., & Schmandt, C. [103]	24	40

Table 2: A Sample of *Polyadic* ACM Papers that Conducted UX Research.

Year	Authors	ACM Citations	Google Scholar Citations
2000	Isbister, K., Nakanishi, H., Ishida, T., & Nass, C. [70]	88	207
2010	Bohus, D., & Horvitz, E. [23]	81	149
2016	Savage, S., Monroy-Hernandez, A., & Höllerer, T. [149]	56	113
2009	Bohus, D., & Horvitz, E. [22]	55	96
2009	Bohus, D., & Horvitz, E. [21]	47	95
2018	Sebo, S. S., Traeger, M., Jung, M., & Scassellati, B. [161]	43	76
2018	Shamekhi, A., Liao, Q. V., Wang D., Bellamy, R. K. E., & Erickson, T. [155]	46	73
2018	Zhang, A. X., & Cranshaw, J. [189]	39	61
2010	Kumar, R., Ai, H., Beuth, J. L., & Rosé, C. P. [85]	27	68
2018	Toxtli, C., Monroy-Hernández, A., & Cranshaw, J. [168]	35	59
2017	Cranshaw, J., Elwany, E., Newman, T., Kocielnik, R., Yu, B. W., Soni, S., Teevan, J., & Monroy-Hernández A. [42]	21	67
2003	Nakanishi, H., Nakazawa, S., Ishida, T., Takanashi, K., & Isbister, K. [118]	9	65
2018	Avula, S., Chadwick, G., Arguello, J., & Capra, R. [9]	20	42
2010	Ai, H., Kumar, R., Nguyen, D., Nagasunder, A., & Rosé, C. P. [5]	3	56
2013	Dyke, G., Howley, I., Adamson, D., Kumar, R., & Rosé, C. P. [48]	1	52
2018	Seering, J., Flores, J. P., Savage, S., & Hammer, J. [152]	23	29

4.3.1 Shared Aspects between Polyadic and Dyadic CAs. We present aspects that *polyadic* and *dyadic* CAs shared, i.e., application domain, modality, agent characteristics, design originality, design method, and evaluation methods below.

Application domain. Among the 36 ACM papers we reviewed, eight *polyadic* CAs have been applied to education and collaborative learning, e.g., [48]; five to online communities, e.g., [149]; three to group discussions, e.g., [77]; seven to work and productivity, e.g., [9, 168]), two to virtual meetings, e.g., [70]; three to guiding services, e.g., [190]; two to games, e.g., [140], one to family, i.e., [102], and five undefined depending on the major focuses of the papers.

Modality. There are 22 polyadic papers applying text-only interactions, e.g., [132]; nine studies are based on video, e.g., [118], only a few are audio only, e.g., [102], and hybrid of audio-text, e.g., [70]. These categories are not mutually exclusive, because some studies implemented multiple designs with different modalities, e.g., [155].

Agent characteristics. We also looked at whether these CAs in the reviewed papers are embodied or not. While 22 papers studied CAs that are not embodied, e.g., [37], the rest of them are embodied CAs, e.g., [70], which have figures that can demonstrate certain social characteristics through animated body languages [32].

Design originality. Most studies present and evaluate original designs, e.g., [37] while three test or adopt existing systems. One paper uses existing systems (i.e., Google Allo) [66] and two papers examine existing bots on the Twitter platform [1, 152].

Design methods. Among those papers which specified their design methods, five of them are framework-based, which means that the designed features are proof-of-concept designs guided by prior frameworks, models, or theories [48, 77, 118, 164, 179]. Meanwhile, most of the designed features are proposed by researchers without leveraging prior designs or catering to specific user-needs,

e.g., [46, 154, 168]. Two CAs adopted participatory design methods, including need-finding interviews, e.g., [76, 189] and two ran ideation workshops, e.g., [12, 102].

Evaluation methods. The most common evaluation method in the reviewed studies is experimental, with 16 papers using it. Other methods include six using surveys, six using interviews, five using field studies, three using Wizard of Oz, two using interaction log analysis, two using observation and one using focus groups. These categories are not mutually exclusive since some studies employed multiple methods in their evaluations.

4.3.2 Unique Aspects with the Polyadic CAs. Below, we present aspects that involve multi-user interaction, unique to *polyadic* CAs, i.e., relationship type and social scale. Moreover, we report related theories and frameworks used in prior *polyadic* works.

Relationship types. The specific relationships highly depend on the contexts of the interactions, which include eight papers designing *polyadic* CAs as co-learners, e.g., [37], four as co-workers, e.g., [42], seven as collaborators/discussants, e.g., collaborators on Slack [9], three as people meeting for the first time, e.g., [70], four as online community members, e.g., [149], three as public visitors and guests, e.g., [190], six as co-players in a game, e.g., [140], one as SMS contacts, e.g., [66], and one as family members, e.g., [102]. Since *polyadic* CAs are designed for multiple users in contrast to *dyadic* CAs for one-on-one interactions, the relationship types between users are a unique dimension for us to examine *polyadic* human-AI interactions.

Social scale. The social scale of the human users is another unique dimension of *polyadic* human-AI interactions since the designs examined in these papers are for two-individuals or multi-users (more than three users). There are 18 papers on multi-users. Combining the relationship types, we can see that the multi-users scale ranges from smaller groups (e.g., three to five co-learners, [48]), to medium sized groups (e.g., ten discussants [85]), and to larger groups (e.g., online community members on Twitter [149]).

Social theories and related theoretical frameworks. We also collected the theories that motivated the designs of the CAs, i.e., presenting theory-inspired or framework-based designs; other papers are less theory-driven. Among them, self-extension theory has been used to discuss users' sense of ownership of a community-owned CA [101]; the Computers Are Social Actors (CASA) paradigm has been used to compare the patterns of human-CA interactions and human-human interactions [140]; message framing theory has been applied to design the bot messages in implementations and evaluations [149]; balance theory has been leveraged to explore the sense of "balance" in the social dynamics in group interaction mediated by CAs [118]; and the Structural Role Theory [18] has been used to identify the roles that the bots play on Twitch [152].

Other papers build on theoretical frameworks, such as: proposing a harmony model of CA mediation using Benne's categorization of functional roles in small group communication [13]; positing Mutual Theory of Mind as the framework to design for natural long-term human-CA interactions [179]; and applying Input-Process-Output models [84] to explain the process of emotional management in teams [12]. A few papers discuss important concepts in multi-user interactions. For example, vulnerability and trust are discussed in teamwork settings to support a design of robot that is intended to build trust in teams; and Barsade's Ripple Effect [10] is introduced to support hypotheses of a machine agent's positive influence on other individuals in a team just as effectively as a human member [161]. However, the theories and theoretical frameworks used in the limited number of existing publications are rather scattered and disorganized.

4.4 Proven Effects of *Polyadic* CAs on Human-Human Interaction (RQ4)

The reviewed articles reported two major sets of effects of *polyadic* CAs: the effects on *dyadic* interaction, emphasizing the effects on one-on-one interaction between human users and the CAs; and the effects on *polyadic* interaction, referring to the impact of the CAs designed to mediate human-human interaction. The effects of the *polyadic* CAs on *polyadic* human-AI interaction include improving the individual users' learning effectiveness [37, 164], perceived self-efficacy [164], and satisfaction [46]. Another large portion of this set of effects involves users' positive or negative attitudes, perceptions, and acceptance toward the CAs [5, 118, 140, 179], and interactions and engagement patterns between users and the CAs depending on specific design features [5, 161]. Most studies laid primary focus on *dyadic* human-AI interaction effects, while the evaluation on group effects is relatively insufficient. Overall, the effects reported in the reviewed papers show opportunities of using *polyadic* CAs to improve human social experience.

4.4.1 Communication Efficiency. In collaborations, studies have supported that *polyadic* CAs can help with consensus-reaching, aid communication comprehension, enhance task management, and save the time and energy of human collaborators from tedious work, which can improve the group work efficiency and overall collaborative experience. Specifically, for consensus-reaching, a moderator CA helps to align group consensus and individual opinions, contributing to reaching agreements [76, 77]. For better comprehension of group communication, a CA can tag and summarize the group chats to help users make sense of the conversation better [189]. Moreover, a tour guide CA in museums can mediate visitors' interactions by prompting topics, offering information, and concluding discussions [124, 125]. In terms of management and coordination, a scheduling CA can coordinate schedules of team members fast and efficiently, setting human personal assistants free [42]. For burdensome tasks, a searchbot is designed for its human collaborators to save time when switching between tools and searching independently by offloading these tasks to the CA.

4.4.2 Group Engagement. The papers we reviewed have provided evidence that *polyadic* CAs may benefit group dynamics through encouraging engagement and balancing uneven participation. The effects of encouraging engagement are most salient for collaborations in education and online communities. In multiple studies in collaborative learning, the intervention of a CA that can ask questions and prompt students to show their thought processes, stimulating pedagogically beneficial conversations in the learning groups [48, 164]. In online communities such as Twitter, bots designed to call people to action to specific social activities engage users to make relevant contributions to the discussion and lead to their interactions regarding future collaborations [149].

Also, *Polyadic* CAs as facilitators in group discussions not only encourage the amount of engagement, but also improve the distribution of the engagement by nudging members to participate evenly [155]. Several CA designs can balance discussion by inviting less engaged learners to join the conversation and downplaying the "talkative" learners from over-controlling conversations [37]. The BabyBot on Twitch can grow up and learn from human users, which engages community members as a whole and opens opportunities for newcomers to participate in interactions as a way to welcome them on board [101, 154].

CAs are also designed to moderate multi-user engagement in open and dynamic environments [21–23], e.g., the hotel reception, games, and organizational systems, by identifying partakers and bystanders, and facilitating engagement decisions of when and whom to engage. Furthermore, more intensive and even engagement brings more diverse content. The moderator CA for discussions can therefore help with generating diverse and deliberative opinions [76, 77].

4.4.3 Relationship Maintenance. Regarding social and relational aspects in groups, *polyadic* CAs show potential in regulating emotions and relationships to maintain harmonized group dynamics. Prior work presented prototype designs of CAs to offset a shortcoming of the text-based communication in teams, that is, the insufficient ability of understanding and managing emotions of the team members [12, 86]. By monitoring the sentiment of the conversations and providing suggestions, CAs can improve emotion regulation and compromise facilitation [12]. Similarly, *polyadic* CA can analyze group behaviors to reinforce positivity by preventing the use of negative words [132]. Moreover, designing an agent that can make vulnerable statements in collaboration generates a *Ripple Effect*, which notes that human teammates with an agent that expresses vulnerability are more likely to engage in trust-related behaviors, including explaining failures to the group, consoling team members who made mistakes, and laughing together. These actions reduce tension in the team and enhance a positive and trusting atmosphere in groups [161].

4.4.4 Building Connections. In a two-individuals interaction, the existence of a CA is crucial to the balance and solidarity of the triad. Studies have shown that a CA's agreement, disagreement, or bias toward one or both side(s) of the two individuals can significantly change the mutual perceptions of the interaction partners [5, 118]. Also, when a CA agrees or disagrees with both the human pair, the interaction partners feel more similar and attractive to each other [118]. In contrast, a CA tutor showing a bias towards one perspective or another will polarize the learning pair [5]. Thus, *polyadic* CAs demonstrate potential to build solidarity between interaction pairs.

Polyadic CAs may also be used to establish new connections between pairs that meet for the first time. i.e., a helper agent that is built to form social connections between people in cross-cultural conversations [70]. The evaluation showed the positive effects of the agent designed to build common ground. Also, a CA was designed as a virtual companion in a university setting to help first-year students get on board and build new connections [99]. However, there were also a series of mixed findings of nuanced human users' social perceptions regarding self, partner, and cultural stereotypes, which are worth further exploration.

4.5 Evaluation Metrics of CAs (RQ5)

We synthesize the metrics of prior studies used to evaluate user experience of conversational agents. These metrics show what categories previous researchers considered in evaluating their designs and what measurements they used. In the following section, we report on the results emerging from the analysis of the 135 *dyadic* papers and the 36 *polyadic* papers included in the final corpus. We categorize them in three main areas: chat log analysis or observation, survey scale, and interview. For definitions of all the metrics, see Appendix A.1 and A.2 for details.

4.5.1 Log Analysis or Observations. For *dyadic* papers, 36 papers evaluated CAs using log analysis or observation methods such as coding user behaviors through watching recorded videos. Table 4. provides an overview of what metrics were evaluated in the user studies. Six categories emerged: 1) linguistic features, i.e., language patterns that are evident in the language use; 2) prosodic features, i.e., features that appear when we put sounds together in connected speech; 3) dialogue features, i.e., features that are related to conversation structure and chat flow; 4) tasks-related features, i.e., features that measures metrics related to task fulfillment; 5) algorithm features, i.e., features concerning model training and efficiency; and 6) user-related features, i.e., features that are closely related to user perceptions and behaviors.

Linguistic features were evaluated in seven recent papers. To analyze the linguistic features in the chat logs, most of these studies leveraged existing natural language processing toolkits to analyze the text. For example, LIWC [74] uses 82 dimensions to determine if a text uses positive or negative emotions, self-references, causal words etc. Another way is to count the use of linguistic patterns, such as the use of pronouns and proportion of utterances with terms repeated from previous conversations [166]. Similar to linguistic features, prosodic features were used to evaluate the language pattern, but especially in speech-based agents, e.g., prior work measured the rate of speech, pitch variation, loudness variation, using OpenSMILE to process the audio signals [166].

Dialogue features were used earlier than linguistic features, and were evaluated in nine papers, measuring the dialogue quality and efficiency [182], e.g., length, dialogue duration, number of turn taking, response time; and dialogue expressiveness, e.g., percentage of sentences. One of the frameworks being widely adopted in prior works is the PARAdigm for dialogue system evaluation, also known as PARADISE framework, which examines user satisfaction based on measures representing the performance dimensions of task success, dialogue quality, and dialogue efficiency, and has been applied to a wide range of systems, e.g., [54].

Task-related features were examined in nine papers. This dimension helps researchers to understand how users fulfill the tasks assigned by or facilitated by CAs. These dimensions are useful for task-oriented CAs that were designed to help users execute a task or solve a problem, e.g., customer service. While the task fulfillment rate was continually used as a metric in evaluating tasks, e.g., [45, 107, 131], some metrics were introduced in recent years,

Table 3: *Polyadic* ACM Papers using Different Metrics for UX Evaluation (Count by Year Since 2005)

Category	Subcategory	Metrics (Defined in Appendix A.1)	(20)05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
Chatlog	Sociability	discussion quality	1				1	1											
		consensus reaching						1											
		opinion expression						1											
		opinion diversity						1											
		even participation						1											
		linguistic features																1	
		message quantity						1								1		1	
	Task-related	task completion time										1							
		efficiency																1	
		effectiveness																1	
		satisfaction																1	
		team performance										1						1	
Survey	Communication-quality	perceived communication effectiveness						1											
		perceived communication fairness						1										1	
		perceived communication efficiency						1											
		perceived group climate														1			
	Socio-emotional	perception of other group members										1	1			1		1	
		perception towards the agent										1							
		perception about collaboration																	
		perceived social presence												1		1			
		perceived social support												1					
		level of annoyance with the intervention																	
		opinion of oneself/partner/cultural stereotypes						1											
		perceived appropriateness of agent's social behavior														1			
		users' psychological wellbeing																1	
	Traditional	anthropomrophism														1			
		perceived emotional competence												1					
		unmet expectations												1					
		privacy concerns												1					
		level of perceived conversation control					1												
		perceived satisfaction						1										1	
		performance/effectiveness														1		1	
		level of engagement																	
Interview	Socialbility	perceived capability to promote contributions														1			
	Issue-relevant	decisions to (not) engage with the CA																1	
		overall UX														1		1	
		reflections on selves and others																1	
		willingness to take actions												1					
		direction of improvement														1		1	

such as dropout rate, e.g., [78], to capture the percentage of all participants who did not complete the task during evaluation.

Eight papers evaluated user perceptions and social behaviors by analyzing the chat log between the CA and the users, using metrics such as self-disclosure [92, 93], intimacy [147], and aggressiveness [25]. Some theoretical frameworks were used, e.g., Barak and Gluck-Ofri's scale, which conceptualized self-disclosure into three dimensions as information, thoughts, and feelings [92].

For *polyadic* papers we reviewed, 14 studies employed log analysis and were mostly published during the 2010s. Several papers used task-oriented metrics, see Table 3, e.g., task duration, efficiency, effectiveness, and team performance [9, 86, 101, 154]. The evaluation metrics were both quantitative and qualitative. The quantitative metrics counted the frequency of users' input or utterances [1, 9, 46, 119, 124, 125, 164]. The qualitative metrics evaluated discussion quality [77, 164, 179], with an emphasis on consensus reaching [77], opinion diversity [164], and linguistic features, e.g., verbosity [179], readability [179], adaptability [179] and positive and negative sentiment [66, 118, 119, 179].

4.5.2 Survey Scale. For *dyadic* papers, 93 papers evaluated CAs using survey scales. Table 5 provides an overview of what features were evaluated in the user studies. Four categories emerged: 1) conversation related metrics: aspects that help CAs to manage dialogues during interactions; 2) user perception of CA's social features: aspects that reflect user perception of CA following social protocols; 3) user perceived system usability: aspects that capture the quality of a UX when users interact with CAs; and 4) user self-reported experience with CAs.

For conversation related features, some studies leverage widely adopted scales because they are often used for system evaluation.

Table 4: *Dyadic* ACM Papers using Chat Analysis for UX Evaluation (Count by Year Since 2005)

Category	Metrics (Defined in Appendix A.2)	(20)05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
Linguistic features	positive/negative words															1	3	
	length of utterances																2	
	personal pronouns																2	
	term-level repetition																1	
	utterance-level repetition																1	
	use of concrete words																1	
	variation in language used																1	
Prosodic features	variation in speech-based agents																1	
Dialogue features	dialogue/response quality					1											3	
	dialogue efficiency metrics			1		1		1									2	
	expressiveness		1															
Tasks-related	task fulfillment rate/effectiveness					1		2							1			
	system usage														3			
	user entry time							1										
	dropout rate															1		
Algorithm	Intent modeling evaluation																1	
	error rate							1										
User related	self-disclosure														1	1	1	1
	intimacy														1			
	reflection														1			
	impolite/aggressive behaviors														1			
	motivation																1	
	affective states																1	

For example, informativeness, or information quality, has been used multiple times in user evaluations. Informativeness refers to quality of the conversational system to communicate truthfully, provide relevant information, and share content clearly and orderly. Prior work on AI-powered CA used Gricean Maxims' information quality metrics [184] to evaluate this. Other metrics were proposed as modes of examination in recent years, e.g., syntactic readability, self-reported response quality and intensity of sentiments[177], and instruction quality [54].

Multiple features were proposed in evaluating users perceptions of conversational agents' social features, such as perceived agents' common ground [35], sociability, [178], and perceived interruption or annoyance [150]). Metrics such as playfulness [183, 192] (enjoyment, interestingness, or funny) and perceived intimacy (perceived interpersonal closeness, or friendliness) have been evaluated using existing scales. For example, the inclusion of others in the self scale and the subjective closeness index to measure level of closeness [187]. With the development of conversational agents, more social features have been introduced in user evaluation, such as perceived anthropomorphism [145], perceived personality traits [172], and perceived naturalness [156].

For system usability evaluation, 33 papers evaluate CA usability using traditional usability metrics. For example, UMUX-LITE is the usability metric for user experience to measure evaluating factors such as perceived ease of use, and it is used in papers [15, 19]. Similarly, NASA-TLX, or task load index, which evaluates mental demand, physical demand, effort, frustration, and future adoption willingness, is used in studies [71, 72]. Moreover, prior work also uses usability questions from previous surveys [75].

There are 45 papers evaluating the conversational agent through collecting metrics regarding user self-reported personal experience responses. Among them, satisfaction, e.g., [15, 45, 97] and perceived trust, e.g., [65, 90, 92, 192], and perceived engagement, e.g., [30, 115, 156, 182] were evaluated the most in the sample. There are some new features been used to evaluate *dyadic* interactions, such as serendipity [30], social orientation toward CAs [8], degree of collaboration [111], and perceived capability to control [166]. Thus, evaluation dimensions are becoming increasingly diverse when examining user experience with conversational agents.

Polyadic studies which employed surveys mainly evaluated users' perceptions of both the functional and socio-emotional aspects of the interactions, with specific focus on communication quality and interpersonal dynamics. The former includes perceived communication effectiveness [77, 164], communication efficiency [77, 101, 164], communication fairness [76, 77], and the quality of collaboration [9]. The latter includes perceived group climate [59, 152], social presence [12, 152], social support [12], perceptions on the agent [46, 86, 152, 154], appropriateness of the agent's intervention [9], as well as impressions about other group members [59, 86, 154]. We observe that during the past decade works evaluated social dynamics in the interactions [12, 59]. These studies with survey methods also used a series of metrics that are commonly used in *dyadic* interactions, which evaluate the ability and competence of CAs [9, 12, 152], task performances [152, 154], perceived anthropomorphism [59, 152, 155], intelligence [155], likeability [59], users' satisfaction of the interaction [77, 154], perceived control of the interaction [118], and level of engagement [9, 152].

4.5.3 Interview. For *dyadic* interaction papers, overall, 17 papers evaluated CAs using interview data analysis. Table 6. provides an overview of what metrics were evaluated in the user studies. On the one hand, users were interviewed regarding their general impressions on the CAs, e.g., [33, 78, 93, 94, 137], perceived CA characteristics such as capability in handling requests [52], personality [78, 137], and trust [89]. On the other hand, they interviewed

Table 5: *Dyadic* ACM Papers using Survey Scales for UX Evaluation (Count by Year Since 2005)

Category	Metrics (Defined in Appendix A.2)	(20)05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21
Conversation related	syntactic readability																1	
	self-reported response quality																1	
	informativeness/information quality	1													1		1	
	conversation smoothness															1		
	intensity of sentiments																1	
	instruction quality					1												
Perception towards CA	perceived common ground										1							
	opinion of the CA as a partner					1												
	liking attitude	1										1			1			
	playfulness/enjoyment	1								1					1	3	3	1
	socialbility														1		1	
	perceived anthropomorphism																2	
	perceived warmth of the AI system																1	
	impression																1	
	perceived personality traits																1	
	perceived persuasiveness					1												
	interaption/annoyance	1																
	perceived intimacy	1												1	1	2	2	1
	perceived naturalness																1	
	perceived safeness															1		
System usability	overall usability									1			1			4	2	2
	perceived ease of use	1	1									1			1		1	
	mental demand														2		1	
	physical demand														2			
	perceived task completion					1					1				2			
	effort														2			
	frustation														2			
	desire to continue the interaction									1					2		2	
	user acceptance of the agent									1					2		2	
	system consistency																1	
	perceived usefulness/helpfulness															1	4	1
	willingness to recommend																1	
	motivation															1		
UX with CA	overall UX														1	1	2	
	contrast of user experience																1	
	perceived quality of the interaction															1	1	
	perceived quality of CA															1	1	
	perceived satisfaction	1		1		1		1		1				1	1	3	3	
	perceived engagement															1	3	2
	perceived trust	1					1								1	1	2	1
	confidence															1		1
	perceived capability to control																1	1
	aroused emotions					1											1	
	serendipity																	1
	pleasant surprise																	1
	empathy															1	1	1
	self-reflection/self-awareness														1		1	1
	self-disclose														1		1	
	anxiety/tension																1	
	comfort																1	
	social orientation toward CAs															1		
	degree of collaboration																	1
	degree of decision-making																1	1

user behaviors towards conversational agents such as efforts used while interacting with the CAs [93], actual engagements [93, 144], and daily using practices [94].

For *polyadic* interaction papers, studies that employed interviews focused more on specific issue-relevant metrics. These metrics include perceived capability of the agent on solving a particular issue [59], engagement in the conversation [9, 119, 154], willingness to take actions under the persuasion of the CA [149], reflections on selves and others [119, 154]. Some general metrics include overall user experience and impression[119, 154, 155], as well as potential directions for improvement[119, 155]. Social interaction features seem to be essential for *polyadic* papers, especially factors related to promoting team contributions and engagement, e.g., [9]. There are many features that have been evaluated in *dyadic* interactions

Table 6: *Dyadic* ACM Papers using Interview Related Metrics for UX Evaluation

Category	Metrics (Defined in Appendix A.2)	Number of Papers (Between 2005 and 2021)
Perceptions	overall impressions and experience	5
	perceived CA's capabilities	1
	perceived CA appearance and self-presentation	1
	perceived personality	2
	perceived naturalness	1
	reciprocative	1
	perceived benefits of the agent	1
	motivation to use CAs	2
	perceived burden	1
	the best and worst aspects of using CAs	1
	perceived human-likeness	1
	perceived communicative experiences	1
	perceived enjoyment	1
	trust	1
	perceived effectiveness	2
Behaviors	engagement	2
	notice of CA's certain function	1
	efforts in learning the system	1
	effects on self-warenss, self-reflection	1
	effect on judgement	1
	effects on collaborations	1
	daily practices of using the CA	1

but are also suitable to be evaluated in the *polyadic* context, such as effect on judgement [11], and perceived burden, i.e., time, financial, mental, and emotional burden [130].

4.6 Overlooked Issues of *Polyadic* CAs (RQ6)

We also reviewed overlooked issues and additional findings aside from the primary design focuses, which raise intriguing issues that can be missing in design guidelines. These issues are related to appropriateness, privacy, and ethics in the designs of CAs, which warrant deeper discussions about the role of *polyadic* CAs, their social influence, and their relationship with human users in different group settings.

4.6.1 Polyadic CAs Need to be "Visible". One overlooked issue points to the users' awareness of the *polyadic* CAs in the group. [9] presented a searchbot to support collaborative search tasks, and the authors discussed that the collaborative nature of their searchbot posed a new issue regarding awareness of the CA. They suggested that the searchbot should announce itself and remain "visible" to the users throughout the interaction. Whether and how to keep users' awareness of *polyadic* CAs were discussed in diverse contexts, such as the tutor [5, 37], the assistant [9, 42], the moderator or facilitator [66, 70, 76, 77], and one of the members in online communities[149, 152]. The findings suggest a direction regarding design decisions to make users aware of CAs as their peers or not.

4.6.2 Polyadic CAs Need to be "Ignorable". There is some discussion of ignorable design in the reviewed papers. In collaborative learning contexts, [164] mentioned that CAs are sometimes ignored and abused by the group learner. Authors found that students provided hasty answers to the tutor agent and sometimes wanted to pay more attention to the learning questions instead of the agent's facilitation. Similarly, participants perceived a task reminder CA to be invasive or annoying, as they are "too frequent", "not context sensitive", and distracting [168].

Only one paper mentioned "easy to ignore" as a design suggestion when developing a supporting agent in multi-user contexts [70], because if CAs are persistent with their intervention, the effects can potentially backfire. Given that there are limited papers discussing ignorable design suggestions, it remains a crucial issue for designers to create CAs that are less intrusive and are able to detect the moment when users do not want their interventions and shut down properly, particularly when the interaction between human users is the major focus.

4.6.3 Polyadic CAs Need to be Accountable. Another overlooked issue arises in voice activated CAs in interpersonal spaces[101]. As CAs are involved in multiple users' interactions in their homes, participants were confused about how an CA would deal with interpersonal conflicts between users in the home without invading privacy[101]. Papers also discussed the issue of CA ownerships - who should CAs be accountable to? If a CA is the mediator between human users, how much can and should other users consider this mediator? If there are interpersonal conflicts, what is the standpoint of the CA? What will happen if the agent crosses a perceived boundary, and how should we tackle it? Several questions such as these were asked in the reviewed papers [101, 154, 191].

4.6.4 Topics Discussed in Polyadic and Dyadic Works. To explore the difference between the topics discussed in the UX research of *dyadic* and *polyadic* CAs, we applied topic modeling method over the papers' *discussion* sections in which authors reflected on the findings and proposed future research directions. As shown in Figure 2, we found that the top topics identified from the *polyadic* CA studies covered different concepts that were not frequent in the *dyadic* CA articles. As shown in Figure 2, *polyadic* papers' Topic 1 was about *group discussion* (18.82% topic weight); whereas *dyadic* papers mentioned *user-agent interaction* the most (44.16% topic weight). For example, Peng et al. evaluated a *polyadic* CA [132] that facilitated emotion regulation in group discussions, which

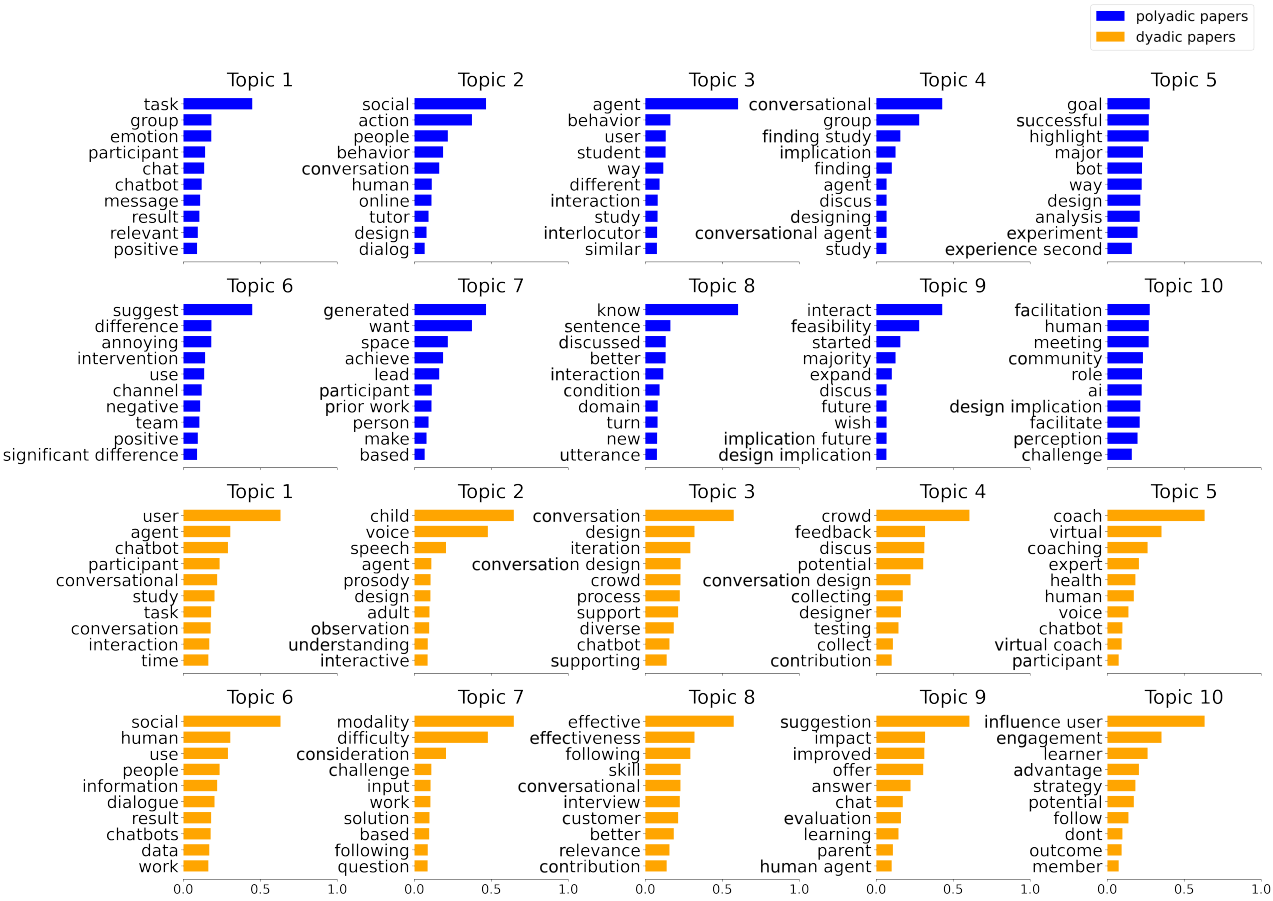

Figure 2: Comparison between topics (ordered by weight) discussed in *polyadic* and *dyadic* papers (the x-axis is the probability distribution of each term in its corresponding topic)

received the highest weight for the *group discussion* topic. This work also contributed to *polyadic* Topic 2 (about *social behavior*, 11.98% topic weight), as it discussed how to positively impact users' social behavior and emotion. A similar topic appeared in the *dyadic* works but was ranked much lower in popularity (Topic 6 with 5.86% weight). On the contrary, Muresan et al.'s paper [115], a *dyadic* CA weighted the highest for Topic 1 *user-agent interaction*, discussed how anthropomorphism of CAs affects users' engagement. This study also contributed to *dyadic* Topic 2 (*conversational design* with 8.08% topic weight). However, both topics were not appeared in the top 10 of *polyadic* works. In addition to *social behavior* (e.g., [86, 140]), *polyadic* CA works tended to discuss more on *education* (Topic 3, 10.94% weight, e.g., [48, 76]), and *embodied design* (Topic 5, 10.13% weight, e.g., [23, 154]). But *education* related topic was also ranked low in *dyadic* papers (Topic 10 with 4.49% weight).

5 DISCUSSION

Given the above results of the literature review, we discuss several research opportunities for future design and development of *polyadic* CAs to address social challenges in human-to-human interactions. These opportunities emerge in three main directions, i.e., exploring an under-studied design space, using and developing theoretical foundations, and developing HCI design guidelines for building boundary-aware CAs. We also discuss specific research agendas of future communicative AI technologies [63] around two key aspects, e.g., relational dynamics and functional dimensions.

5.1 Spotlight an Under-Explored Design Space

Overall, we found a very small fraction of CAs, published in ACM venues, designed for *polyadic* human-AI interaction. Results suggest that such a design space has been severely under-explored. A predominance of CA surveys worked on the classification of CAs [32, 44, 53, 69, 79, 88, 153], and some efforts were made in investigating CAs' potential to improve user engagement and experience in *dyadic* Human-AI interaction [38, 139]. However, still little attention was given to the conversational effects of CAs on *polyadic* human-AI interaction. Thus, this literature review filled the research gap and mapped the progress in this field for future researchers.

In terms of the application domains, results showed that some areas (e.g., public services, health) fall short of UX research on *polyadic* CAs. Also, user evaluation lacks empirical findings of large

scaled human-human interaction. We also found that most original designs were proposed to explore the feasibility of conversational AI in varied application domains (e.g., [12, 46, 76, 102, 154, 168]. With the tsunami of conversational AI, the need for design knowledge becomes more intense [139]. However, there is no wide consensus on grounded practices on which to create novel designs.

5.2 Connect with and Contribute to Theories

Our review highlights a limitation in the theoretical groundings of designing and evaluating CAs for *polyadic* Human-AI interactions. We found that many works did not leverage or contribute to existing theories. This was also found by prior literature reviews of CAs [41, 139]. There can be two reasons. First, *polyadic* conversational design is still an emerging area, and HCI researchers have not identified the best theories to explain, build on, and support their studies. Currently, some adopted existing theoretical grounds from social psychology (e.g., CASA) [140], sociology (e.g., balance theory, structural roles) [118, 152] and communication studies (e.g., self-extension theory) [101] . Second, *polyadic* CAs are bringing brand-new dynamics of social interactions, to an extent that existing theoretical frameworks cannot account for. For example, prior work [55] identified that human users' expectations and perceptions towards machines are different from those towards human partners. Thus, when new social dynamics are formed, existing theories may not be sufficient.

To leverage existing theories, scholars need to identify the social interaction *challenges* (e.g., findings in RQ1) to be addressed and the *application domains* (e.g., findings in RQ3), and then look for related theories resolving the particular challenge in the relevant application domain to inspire the CA designs. For example, some *polyadic* CAs are designed for group members to reach consensus [12, 76, 77]. Prior works on collaborative engineering, a domain where designs are created to enable engineers to work effectively with all stakeholders for completing collaborative tasks [26], have applied consensus building theory (CBT) [28] in designing for collaboration processes [82, 117]. The CBT is often applied by those conducting IT requirement negotiations or those conducting risk and control self-assessments. In the situation where decisions cannot be made unilaterally because team members are co-responsible peers, teams need to resort to approaches to build consensus to gain commitment [82]. The CBT could be applied to direct four major steps in reaching a consensus, i.e., articulating a proposal, evaluating willingness to commit, diagnosing causes of conflict, and invoking conflict resolution strategies [28]. Similarly, when designing a *polyadic* CA under a collaborative negotiation situation, designers may leverage the CBT model to inform designs. Future studies can also adopt a similar approach in finding suitable theories or frameworks.

5.3 Propose the Notion of Boundary-Awareness

Only a few *polyadic* papers used existing systems in user evaluation. Regardless of the CAs' originality, the overlooked negative UX identified by prior research indicates that there is a pressing need for improvements in design strategies and guidelines.

5.3.1 Polyadic CAs are Unique to Design. In prior works, human-likeness (e.g., empathy [146] and self-disclosure [94]) is identified as a critical aspect to improving UX and to encouraging users to show favorable feelings (trust, openness, tolerance, etc) towards CAs [38, 139]. However, similar topics are less reflected by *polyadic* human-AI interaction (RQ5). Instead, researchers raised most open concerns on how CAs should establish social boundaries in human-human interactions, which suggests that *polyadic* CAs may be *unique* to design compared with *dyadic* CAs.

Taking the learning context as an example, when a *polyadic* CA was deployed as a study peer, researchers found that students teams often ignore or provide hasty replies to CAs [164]. However, how can CAs be less intrusive partners in such a context? Similarly, in a family context, when family members have different opinions, how can the CA understand when to "knock the door" when it is tackling conflicts between couples [191]? Should it let parents know where their children are up to [102]? Our findings highlight the importance of questions for CAs to understand social boundaries. Current studies are yet able to take care of the dilemmas in situations where user needs or human-AI interaction conflict and the CAs need to react or take sides.

5.3.2 Design CAs with Boundary-Awareness. By reviewing the overlooked issues, we identified multi-dimensional themes in *polyadic* human-AI interaction, which we consider as social boundary issues (RQ6). Boundaries in social sciences can be understood as a set of rules followed by most people in a particular society, which are vital in the society because they can guide human behaviors and assist in managing what is and what is not acceptable [68, 87]. This issue is closely related to privacy and disclosure, as the presence of *polyadic* CAs as social actors changes the social dynamics and users' information management strategies.

Communication privacy management theory (CPM) [133] identified three main elements in people's private information management: (1) privacy ownership, (2) privacy control, and (3) privacy turbulence. When an individual has decided to disclose private information, as a result, the recipient of the information becomes a co-owner or shareholder. From that moment, the initial owner of the information must set rules and boundaries of how to manage private information. Privacy turbulence may happen when these rules are violated, and the original owner may refrain from disclosing any more information or may engage in negotiations or coordinate rules and boundaries with the co-owner [134]. Clarifying ownership of users' private data to *polyadic* CAs and enabling them to learn privacy boundaries in relationships over time are essential steps towards building social sensibility *polyadic* CAs when participating in human-human interactions. For example, a boundary-aware CA can learn if and when it should intervene when mediating couples' conflicts by receiving requests from one side of the couple [191]. Similarly, when users request a CA to understand situations of their living alone elderly parents, the boundary aware CA can reply appropriately. The privacy boundaries in the interactions also depend on how CAs present themselves and how users perceive the CAs. For instance, when CAs join human interactions, their roles may vary from active conversation participants to less salient group collaboration assistants. A boundary-aware CA may adjust its proactivity and intensity in joining the discussion.

In our review, multiples studies showed that during the interactions, users expressed confusions and concerns about what roles

polyadic CAs should play, how *polyadic* CAs should behave and what is the proper distance that *polyadic* CAs should maintain with their users in various group settings. As *polyadic* human-AI interaction involves multiple relationships and complex social dynamics in its nature [164], the boundary between CAs and different users, and the boundary between users need to be taken care of.

Thus, we propose to design CAs with *boundary-awareness* for supporting *polyadic* human-AI interaction. Traditional HCI design addressed boundaries, e.g., between computers and people and between computers and the physical world [160, 171]. Different from the traditional boundaries, which are often between virtual and physical worlds [29], the new boundaries brought by the CAs involve human-to-human boundaries. Prior work [114] suggested that, when facing AI, humans demonstrate different personalities from interacting with other humans. Thus, it might be difficult for CAs to add boundary management in the design because of the collective result of unclear roles, undecided social rules, and social emotions [62] during human-AI interaction.

CAs with *boundary-awareness* cannot be detached from the discussion of privacy, e.g., [9, 102, 141]. To address the boundary issue such as "whether the CA should inform the parents where their kids are up to" [102], we can gain insights from privacy boundaries. For example, Palen and Dourish [126] discussed three boundaries with different desires during one's privacy management. 1) "Disclosure boundary" is the desire to keep one's information private or public. Privacy regulation in practice is not simply a matter of prohibiting one's data from being disclosed. Instead, human-to-human interactions frequently require selective disclosures of personal information to declare allegiance or even differentiate ourselves from others. 2) "Temporal boundary" suggested that privacy management is in the tension between past, present, and future. Users' response to disclosure situations is interpreted according to other events and expectations. 3) "Identity boundary" implied a boundary between self and other. For example, employees are discouraged from using corporate email addresses to post to public forums. Designers can leverage these three aspects to understand the most critical privacy boundaries in creating CAs with boundary awareness.

5.3.3 Adopt and Evaluate the Existing Guidelines. The unique challenges urge us to re-examine existing AI design guidelines and design notions. For example, Microsoft's guidelines for designing human-AI interaction suggested that AI-infused systems should "support efficient dismissal" [6], therefore CAs should be able to learn the social boundaries such that their interaction is perceived less intrusive, addressing a concern that was raised in prior *polydic* CA works [14, 164]. The guidelines also suggested that AI systems should "match relevant social norms" (i.e., "the experience is delivered in a way that users would expect") and should "mitigate social biases" (i.e., "the system language and behaviors do not reinforce undesirable and unfair stereotypes and biases given their social and cultural context") [6]. An example of potentially using these guidelines can be found in prior *polydic* CA works [77] which designed a CA tailored for structural group discussions. The study reported that the CA facilitates the team reaching a logical consensus on a highly contentious topic even though the members' positions and understandings are vehemently opposed to each other, indicating the design of CA's that acts in a way that the team expected (knowing the norm). However, the authors also proposed that when discussing divisive issues, it may be more appropriate to design with interpersonal and social power dynamics in mind and encourage and protect participants' contributions from marginalized groups (understanding social biases.)

Moreover, existing CA guidelines also provide a broad set of items for designing responsible and trustworthy CAs, such as transparency, human control, awareness of human values, accountability, fairness, privacy and security, accessibility, and professional responsibility, e.g., [109]. These guidelines can be employed in future CA designs, and more academic works should also evaluate these guidelines' effectiveness in the wild.

5.4 Other Key Aspects of Communicative AI

We also discuss research agendas of future communicative AI technologies from two perspectives, i.e., relational dynamics and functional dimensions.

5.4.1 Relational Dynamics. Findings from RQ4 suggest that *polyadic* CAs show potentials for influencing social dynamics. Relational dynamics reflect the ways people interact with communicative technologies, with themselves, and with other people or group of people [63]. Possible research directions include the dynamics of relationship types, and relational attributions. We suggest that *polyadic* CAs can be further investigated in its effectiveness to foster new relational dynamics. For example, according to the Computers as Social Actors paradigm [120], CAs are considered as "social actors". Humans rely on many perceivable attributes during an social interactions with other humans [47]. To make sense of human-AI interaction, prior works identify, implement, and test CA effects in various social attributions, such as "social cue" [49, 123] (i.e., a signal that triggers a social reaction of the user towards the CA), "social roles" [152] (i.e., the function CAs and humans play in the interaction context), and social identities [110, 176] (i.e., how CA or humans organize themselves into and within groups, social values that take collective goals, ethical concerns into consecrations), to name a few.

5.4.2 Functional Dimensions. We propose that the design and use of *polyadic* CAs should address questions such as: 1) what communication challenges are to be addressed by *polyadic* CAs?; 2) what is the unique function of *polyadic* CAs?; and 3) how can *polyadic* CA effectively address the identified challenge compared to other methods? These questions should fit into the functional dimensions proposed by [63], which reflects how certain AI technologies are designed and how people can make sense of these devices and applications.

In our literature review, not all works identified a fundamental challenge of human-human interaction before proposing a *polyadic* CA (RQ1). There could be a tendency of technology-determinism", believing that "technological change can determine social change in a prescribed manner[43]." Namely, some works assumed that CAs could help with the challenges without evaluating the effect of the proposed CAs by comparing with other counterparts, e.g., without a CA, with a human, or with previously tested CAs. In future research, designers and practitioners may be better aware

of this potential tendency and propose more comprehensive study plans to evaluate the proposed CAs.

6 LIMITATIONS AND FUTURE WORK

This literature review has several limitations as a result of the data source, the search query, and data analysis methods. First, we chose ACM DL as the source of our literature review work. Several prior works also conducted literature reviews by using only ACM publications for HCI and CSCW [136, 174] and computer science research [175]. However, using one major source could potentially generate a selection bias, e.g., including certain publications from non-ACM venues into the analysis could be challenged as "distorting the conclusions toward a particular direction" [175]. Therefore, when presenting our findings, we focused on the selected papers as samples illustrating the emerging themes that were not presented in prior literature reviews. One major purpose of doing so is to inform UX researchers and practitioners to design conversational AI by leveraging the existing UX research outcomes and apply the suite of measurements in their work. In the future, conducting a systematic literature review by sampling from different sets of HCI research, e.g., papers indexed by IEEE, Web of Science, Scopus, and Science Direct, can further deepen our understanding of conversational AI research.

Second, we selected research publications that completed both design and UX evaluation of CAs. According to quantitative survey of experimental evaluation in computer science, scholarly publications in ACM could be classified into five categories, including formal theory, design and modeling, empirical work, hypothesis testing, and others (e.g., surveys) [167]. The publications we selected could be a combination of design and modeling (i.e., "systems, techniques, or models, whose claimed properties cannot be proven formally") and empirical work (i.e., "articles that collect, analyze, and interpret observations about known designs, systems, or models, or about abstract theories or subjects [174]"). Even though we tuned the search query to the best we could during web-crawling, it is possible that certain works were not captured by the terms we defined due to our fields of expertise. Meanwhile, it is expected that a significant amount of works that proposed novel techniques without UX evaluations were not included in our analysis. Therefore, we do not claim that the findings are exhaustive. Even though our findings build upon significant prior literature review results, the themes of *polyadic* challenges and evaluation metrics can very and will expand with more works identified.

Lastly, the thematic analysis process is subjective, and lemmatization [121] and topic modeling [61] might introduce potential biases as well. For example, lemmatization reduces words to their lemma forms, the use of lemmatization could potentially cause a loss of words' meaning at a certain level. For topic modeling, since we used text from the original papers without filtering irrelevant words, the results contained noises and irrelevant words that sometimes interfered the interpretability. Therefore, we cannot claim differences to be statistically significant. For example, the topic modeling results were discussed in the context of specific examples to elicit more discussions and future research.

7 CONCLUSION

This literature review presents what fundamental human-human interaction challenges are addressed by *polyadic* CA works scrutinized from ACM publications, what effects on human-human interaction are evaluated by these works, and what issues are overlooked when designing *polyadic* CAs. In particular, we propose that researchers and practitioners should design CAs with *boundary awareness* for supporting *polyadic* human-AI interaction. Further, we envision several research opportunities on conversational human-AI interaction with respect to the theoretical groundings, relational dynamics, and functional dimensions.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 2119589. The research is also partially supported by the IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR), a research collaboration as part of the IBM AI Horizons Network.

REFERENCES

[1] Norah Abokhodair, Daisy Yoo, and David W McDonald. 2015. Dissecting a social botnet: Growth, content and influence in Twitter. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 839–851.

[2] Martin Adam, Michael Wessel, and Alexander Benlian. 2020. AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets* (2020), 1–19.

[3] Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications* 2 (2020), 100006.

[4] Eleni Adamopoulou and Lefteris Moussiades. 2020. An overview of chatbot technology. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 373–383.

[5] Hua Ai, Rohit Kumar, Dong Nguyen, Amrut Nagasunder, and Carolyn P Rosé. 2010. Exploring the effectiveness of social capabilities and goal alignment in computer supported collaborative learning. In *International Conference on Intelligent Tutoring Systems*. Springer, 134–143.

[6] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[7] Mahoro Anabuki, Hiroyuki Kakuta, Hiroyuki Yamamoto, and Hideyuki Tamura. 2000. Welbo: An embodied conversational agent living in mixed reality space. In *CHI'00 extended abstracts on Human factors in computing systems*. 10–11.

[8] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. 2019. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[9] Sandeep Avula, Gordon Chadwick, Jaime Arguello, and Robert Capra. 2018. Searchbots: User engagement with chatbots during collaborative search. In *Proceedings of the 2018 conference on human information interaction & retrieval*. 52–61.

[10] Sigal G Barsade. 2002. The ripple effect: Emotional contagion and its influence on group behavior. *Administrative science quarterly* 47, 4 (2002), 644–675.

[11] Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do Multilingual Users Prefer Chat-bots that Code-mix? Let's Nudge and Find Out! *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

[12] Ivo Benke, Michael Thomas Knierim, and Alexander Maedche. 2020. Chatbot-based emotion management for distributed teams: A participatory design study. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–30.

[13] Kenneth D Benne and Paul Sheats. 1948. Functional roles of group members. *Journal of social issues* 4, 2 (1948), 41–49.

[14] JL Beuth, CP Rosé, and R Kumar. 2010. Software agent-monitored tutorials enabling collaborative learning in computer-aided design and analysis. In *ASME International Mechanical Engineering Congress and Exposition*, Vol. 44434. 339–346.

[15] Timothy W Bickmore, Lisa Caruso, and Kerri Clough-Gorr. 2005. Acceptance and usability of a relational agent interface by urban older adults. In *CHI'05 extended abstracts on Human factors in computing systems*. 1212–1215.

[16] Timothy W Bickmore, Suzanne E Mitchell, Brian W Jack, Michael K Paasche-Orlow, Laura M Pfeifer, and Julie O'Donnell. 2010. Response to a relational agent by hospital patients with depressive symptoms. *Interacting with computers* 22, 4 (2010), 289–298.

[17] Timothy W Bickmore, Laura M Pfeifer, and Brian W Jack. 2009. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1265–1274.

[18] Bruce J Biddle. 1986. Recent developments in role theory. *Annual review of sociology* 12, 1 (1986), 67–92.

[19] Pradipta Biswas. 2006. A flexible approach to natural language generation for disabled children. In *Proceedings of the COLING/ACL 2006 Student Research Workshop*. 1–6.

[20] Eva Bittner, Sarah Oeste-Reiß, and Jan Marco Leimeister. 2019. Where is the bot in our team? Toward a taxonomy of design option combinations for conversational agents in collaborative work. In *Proceedings of the 52nd Hawaii international conference on system sciences*.

[21] Dan Bohus and Eric Horvitz. 2009. Dialog in the open world: platform and applications. In *Proceedings of the 2009 international conference on Multimodal interfaces*. 31–38.

[22] Dan Bohus and Eric Horvitz. 2009. Learning to predict engagement with a spoken dialog system in open-world settings. In *Proceedings of the SIGDIAL 2009 Conference*. 244–252.

[23] Dan Bohus and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. 1–8.

[24] Dan Bohus and Eric Horvitz. 2011. Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings of the SIGDIAL 2011 Conference*. 98–109.

[25] Michael Bonfert, Maximilian Spliethöver, Roman Arzaroli, Marvin Lange, Martin Hanci, and Robert Porzel. 2018. If you ask nicely: a digital assistant rebuking impolite voice commands. In *proceedings of the 20th ACM international conference on multimodal interaction*. 95–102.

[26] Milton Borsato and Margherita Peruzzini. 2015. Collaborative engineering. In *Concurrent engineering in the 21st century*. Springer, 165–196.

[27] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[28] Robert O Briggs, Gwendolyn L Kolfschoten, and Gert-Jan de Vreede. 2005. Toward a theoretical model of consensus building. *AMCIS 2005 Proceedings* (2005), 12.

[29] Kevin Burden and Matthew Kearney. 2016. Conceptualising authentic mobile learning. In *Mobile learning design*. Springer, 27–42.

[30] Wanling Cai, Yucheng Jin, and Li Chen. 2021. Critiquing for Music Exploration in Conversational Recommender Systems. In *26th International Conference on Intelligent User Interfaces*. 480–490.

[31] Jacky Casas, Marc-Olivier Tricot, Omar Abou Khaled, Elena Mugellini, and Philippe Cudré-Mauroux. 2020. Trends & Methods in Chatbot Evaluation. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 280–286.

[32] Justine Cassell. 2000. Embodied conversational interface agents. *Commun. ACM* 43, 4 (2000), 70–78.

[33] Luis Cavazos Quero, Jorge Iranzo Bartolomé, Dongmyeong Lee, Yerin Lee, Sangwon Lee, and Jundong Cho. 2019. Jido: a conversational tactile map for blind people. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 682–684.

[34] Joyce Chai, Veronika Horvath, Nanda Kambhatla, Nicolas Nicolov, and Margo Stys-Budzikowska. 2001. A conversational interface for online shopping. In *Proceedings of the First International Conference on Human Language Technology Research*.

[35] Joyce Y Chai, Lanbo She, Rui Fang, Spencer Ottarson, Cody Littley, Changsong Liu, and Kenneth Hanson. 2014. Collaborative effort towards common ground in situated human-robot dialogue. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 33–40.

[36] Huilin Chang, Yihnew Eshetu, and Celeste Lemrow. 2021. Supervised Machine Learning and Deep Learning Classification Techniques to Identify Scholarly and Research Content. In *2021 Systems and Information Engineering Design Symposium (SIEDS)*. IEEE, 1–6.

[37] Sourish Chaudhuri, Rohit Kumar, Iris K Howley, and Carolyn Penstein Rosé. 2009. Engaging Collaborative Learners with Helping Agents.. In *AIED*. 365–372.

[38] Ana Paula Chaves and Marco Aurelio Gerosa. 2020. How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction* (2020), 1–30.

[39] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758.

[40] Piotr Chynał, Julia Falkowska, and Janusz Sobecki. 2018. Human-Human Interaction: A Neglected Field of Study?. In *International Conference on Intelligent Human Systems Integration*. Springer, 346–351.

[41] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, et al. 2019. The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers* 31, 4 (2019), 349–371.

[42] Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 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*. 2382–2393.

[43] Allan Dafoe. 2015. On technological determinism: a typology, scope conditions, and a mechanism. *Science, Technology, & Human Values* 40, 6 (2015), 1047–1076.

[44] Allan de Barcelos Silva, Marcio Miguel Gomes, Cristiano André da Costa, Rodrigo da Rosa Righi, Jorge Luis Victoria Barbosa, Gustavo Pessin, Geert De Doncker, and Gustavo Federizzi. 2020. Intelligent personal assistants: A systematic literature review. *Expert Systems with Applications* 147 (2020), 113193.

[45] Vera Demberg, Andi Winterboer, and Johanna D Moore. 2011. A strategy for information presentation in spoken dialog systems. *Computational Linguistics* 37, 3 (2011), 489–539.

[46] Kohji Dohsaka, Ryota Asai, Ryuichiro Higashinaka, Yasuhiro Minami, and Eisaku Maeda. 2009. Effects of conversational agents on human communication in thought-evoking multi-party dialogues. In *Proceedings of the SIGDIAL 2009 Conference*. 217–224.

[47] Judith Donath. 2007. Signals, cues and meaning. *Signals, Truth and Design* (2007).

[48] Gregory Dyke, Iris Howley, David Adamson, Rohit Kumar, and Carolyn Penstein Rosé. 2013. Towards academically productive talk supported by conversational agents. In *Productive multivocality in the analysis of group interactions*. Springer, 459–476.

[49] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies* 132 (2019), 138–161.

[50] Asbjørn Følstad, Theo Araujo, Effie Lai-Chong Law, Petter Bae Brandtzaeg, Symeon Papadopoulos, Lea Reis, Marcos Baez, Guy Laban, Patrick McAllister, Carolin Ischen, et al. 2021. Future directions for chatbot research: an interdisciplinary research agenda. *Computing* (2021), 1–28.

[51] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. *interactions* 24, 4 (2017), 38–42.

[52] Asbjørn Følstad and Marita Skjuve. 2019. Chatbots for customer service: user experience and motivation. In *Proceedings of the 1st international conference on conversational user interfaces*. 1–9.

[53] Asbjørn Følstad, Marita Skjuve, and Petter Bae Brandtzaeg. 2018. Different chatbots for different purposes: towards a typology of chatbots to understand interaction design. In *International Conference on Internet Science*. Springer, 145–156.

[54] Mary Ellen Foster, Manuel Giuliani, and Alois Knoll. 2009. Comparing objective and subjective measures of usability in a human-robot dialogue system. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*.

[55] Jesse Fox and Andrew Gambino. 2021. Relationship Development with Humanoid Social Robots: Applying Interpersonal Theories to Human/Robot Interaction. *Cyberpsychology, Behavior, and Social Networking* (2021).

[56] Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1371–1374.

[57] David Griol, Javier Carbó, and José M Molina. 2013. An automatic dialog simulation technique to develop and evaluate interactive conversational agents. *Applied Artificial Intelligence* 27, 9 (2013), 759–780.

[58] Jorne Grolleman, Betsy van Dijk, Anton Nijholt, and Andrée van Emst. 2006. Break the habit! designing an e-therapy intervention using a virtual coach in aid of smoking cessation. In *International Conference on Persuasive Technology*. Springer, 133–141.

[59] Agneta Gulz, Magnus Haake, and Annika Silvervarg. 2011. Extending a teachable agent with a social conversation module–effects on student experiences and learning. In *International conference on artificial intelligence in education*. Springer, 106–114.

[60] Fatih Gurcan and Nergiz Ercil Cagiltay. 2020. Research trends on distance learning: a text mining-based literature review from 2008 to 2018. *Interactive Learning Environments* (2020), 1–22.

[61] Fatih Gurcan, Nergiz Ercil Cagiltay, and Kursat Cagiltay. 2021. Mapping human–computer interaction research themes and trends from its existence to today: A topic modeling-based review of past 60 years. *International Journal of Human–Computer Interaction* 37, 3 (2021), 267–280.

[62] Andrea L Guzman. 2020. Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication* 1, 1 (2020), 3.

[63] Andrea L Guzman and Seth C Lewis. 2020. Artificial intelligence and communication: A Human–Machine Communication research agenda. *New Media & Society* 22, 1 (2020), 70–86.

[64] Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

[65] Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2019. An end-to-end conversational style matching agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. 111–118.

[66] Jess Hohenstein and Malte Jung. 2018. AI-Supported Messaging: An Investigation of Human-Human Text Conversation with AI Support. In *Proceedings of the Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.

[67] Jess Hohenstein and Malte Jung. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106 (2020), 106190.

[68] David J Houghton and Adam N Joinson. 2010. Privacy, social network sites, and social relations. *Journal of Technology in Human Services* 28, 1-2 (2010), 74–94.

[69] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. 2019. A survey on conversational agents/chatbots classification and design techniques. In *Workshops of the International Conference on Advanced Information Networking and Applications*. Springer, 946–956.

[70] Katherine Isbister, Hideyuki Nakanishi, Toru Ishida, and Cliff Nass. 2000. Helper agent: Designing an assistant for human-human interaction in a virtual meeting space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 57–64.

[71] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N Patel. 2018. Convey: Exploring the use of a context view for chatbots. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–6.

[72] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. 2018. Evaluating and informing the design of chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 895–906.

[73] Ruikai Liu Jorj X. McKie. 2016. PyMuPDF. https://https://github.com/pymupdf/PyMuPDF.

[74] Masamune Kawasaki, Naomi Yamashita, Yi-Chieh Lee, and Kayoko Nohara. 2020. Assessing users' mental status from their journaling behavior through chatbots. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[75] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[76] Soomin Kim, Jinsu Eun, Changhoon Oh, Bongwon Suh, and Joonhwan Lee. 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*. 1–13.

[77] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.

[78] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing data from chatbot and web surveys: Effects of platform and conversational style on survey response quality. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[79] Lorenz Cuno Klopfenstein, Saverio Delpriori, Silvia Malatini, and Alessandro Bogliolo. 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.

[80] Ahmet Baki Kocaballi, Shlomo Berkovsky, Juan C Quiroz, Liliana Laranjo, Huong Ly Tong, Dana Rezazadegan, Agustina Briatore, and Enrico Coiera. 2019. The personalization of conversational agents in health care: systematic review. *Journal of medical Internet research* 21, 11 (2019), e15360.

[81] Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. 2018. Designing for workplace reflection: a chat and voice-based conversational agent. In *Proceedings of the 2018 designing interactive systems conference*. 881–894.

[82] Gwendolyn L Kolfschoten and Gert-Jan De Vreede. 2007. The collaboration engineering approach for designing collaboration processes. In *International Conference on Collaboration and Technology*. Springer, 95–110.

[83] Stefan Kopp, Lars Gesellensetter, Nicole C Krämer, and Ipke Wachsmuth. 2005. A conversational agent as museum guide–design and evaluation of a real-world application. In *International workshop on intelligent virtual agents*. Springer, 329–343.

[84] Steve WJ Kozlowski and Daniel R Ilgen. 2006. Enhancing the effectiveness of work groups and teams. *Psychological science in the public interest* 7, 3 (2006), 77–124.

[85] Rohit Kumar, Hua Ai, Jack L Beuth, and Carolyn P Rosé. 2010. Socially capable conversational tutors can be effective in collaborative learning situations. In *International conference on intelligent tutoring systems*. Springer, 156–164.

[86] Rohit Kumar and Carolyn P Rosé. 2014. Triggering effective social support for online groups. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 3, 4 (2014), 1–32.

[87] Michèle Lamont and Virág Molnár. 2002. The study of boundaries in the social sciences. *Annual review of sociology* 28, 1 (2002), 167–195.

[88] Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association* 25, 9 (2018), 1248–1258.

[89] David R Large, Leigh Clark, Gary Burnett, Kyle Harrington, Jacob Luton, Peter Thomas, and Pete Bennett. 2019. " It's small talk, jim, but not as we know it." engendering trust through human-agent conversation in an autonomous, self-driving car. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. 1–7.

[90] Tessa Lau, Julian Cerruti, Guillermo Manzato, Mateo Bengualid, Jeffrey P Bigham, and Jeffrey Nichols. 2010. A conversational interface to web automation. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 229–238.

[91] Minha Lee, Sander Ackermans, Nena van As, Hanwen Chang, Enzo Lucas, and Wijnand IJsselsteijn. 2019. Caring for Vincent: a chatbot for self-compassion. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[92] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2020. Designing a chatbot as a mediator for promoting deep self-disclosure to a real mental health professional. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–27.

[93] Yi-Chieh Lee, Naomi Yamashita, and Yun Huang. 2021. Exploring the Effects of Incorporating Human Experts to Deliver Journaling Guidance through a Chatbot. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–27.

[94] Yi-Chieh Lee, Naomi Yamashita, Yun Huang, and Wai Fu. 2020. " 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.

[95] Q Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N Sadat Shami, and Werner Geyer. 2018. All work and no play?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[96] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)* 4, 4 (2013), 1–28.

[97] Mingming Liu, Qicheng Ding, Yu Zhang, Guoguang Zhao, Changjian Hu, Jiangtao Gong, Penghui Xu, Yu Zhang, Liuxin Zhang, and Qianying Wang. 2020. Cold Comfort Matters-How Channel-Wise Emotional Strategies Help in a Customer Service Chatbot. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–7.

[98] Silvia B Lovato, Anne Marie Piper, and Ellen A Wartella. 2019. Hey Google, do unicorns exist? Conversational agents as a path to answers to children's questions. In *Proceedings of the 18th ACM International Conference on Interaction Design and Children*. 301–313.

[99] Christian Löw and Lukas Moshuber. 2020. Grätzelbot-Gamifying Onboarding to Support Community-Building among University Freshmen. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. 1–3.

[100] Anders Lundkvist and Ali Yakhlef. 2004. Customer involvement in new service development: a conversational approach. *Managing Service Quality: An International Journal* (2004).

[101] Michal Luria, Joseph Seering, Jodi Forlizzi, and John Zimmerman. 2020. Designing Chatbots as Community-Owned Agents. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–3.

[102] Michal Luria, Rebecca Zheng, Bennett Huffman, Shuangni Huang, John Zimmerman, and Jodi Forlizzi. 2020. Social Boundaries for Personal Agents in the Interpersonal Space of the Home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[103] Stefan Marti and Chris Schmandt. 2005. Physical embodiments for mobile communication agents. In *Proceedings of the 18th annual ACM symposium on User interface software and technology*. 231–240.

[104] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. 2019. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[105] Michael McTear. 2020. Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots. *Synthesis Lectures on Human Language Technologies* 13, 3 (2020), 1–251.

[106] Michael Frederick McTear, Zoraida Callejas, and David Griol. 2016. *The conversational interface*. Vol. 6. Springer.

[107] Indrani Medhi, Somani Patnaik, Emma Brunskill, SN Nagasena Gautama, William Thies, and Kentaro Toyama. 2011. Designing mobile interfaces for

novice and low-literacy users. *ACM Transactions on Computer-Human Interaction (TOCHI)* 18, 1 (2011), 1–28.

[108] Eleonora Mencarini, Amon Rapp, Lia Tirabeni, and Massimo Zancanaro. 2019. Designing wearable systems for sports: A review of trends and opportunities in human–computer interaction. *IEEE Transactions on Human-Machine Systems* 49, 4 (2019), 314–325.

[109] Microsoft. 2018. Responsible bots: 10 guidelines for developers of conversational AI. (2018).

[110] Milad Mirbabaie, Stefan Stieglitz, Felix Brünker, Lennart Hofeditz, Björn Ross, and Nicholas RJ Frick. 2021. Understanding collaboration with virtual assistants–the role of social identity and the extended self. *Business & Information Systems Engineering* 63, 1 (2021), 21–37.

[111] Elliot G Mitchell, Rosa Maimone, Andrea Cassells, Jonathan N Tobin, Patricia Davidson, Arlene M Smaldone, and Lena Mamykina. 2021. Automated vs. Human Health Coaching: Exploring Participant and Practitioner Experiences. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–37.

[112] Pragnesh Jay Modi, Manuela Veloso, Stephen F Smith, and Jean Oh. 2004. Cmradar: A personal assistant agent for calendar management. In *International Bi-Conference Workshop on Agent-Oriented Information Systems*. Springer, 169–181.

[113] Joao Luis Zeni Montenegro, Cristiano André da Costa, and Rodrigo da Rosa Righi. 2019. Survey of conversational agents in health. *Expert Systems with Applications* 129 (2019), 56–67.

[114] Yi Mou and Kun Xu. 2017. The media inequality: Comparing the initial human-human and human-AI social interactions. *Computers in Human Behavior* 72 (2017), 432–440.

[115] Andreea Muresan and Henning Pohl. 2019. Chats with bots: Balancing imitation and engagement. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[116] Morten Johan Mygland, Morten Schibbye, Ilias O Pappas, and Polyxeni Vassilakopoulou. 2021. Affordances in human-chatbot interaction: a review of the literature. In *Conference on e-Business, e-Services and e-Society*. Springer, 3–17.

[117] Josephine Nabukenya, Patrick van Bommel, and Henderik Alex Proper. 2009. A theory-driven design approach to collaborative policy making processes. In *2009 42nd Hawaii International Conference on System Sciences*. IEEE, 1–10.

[118] Hideyuki Nakanishi, Satoshi Nakazawa, Toru Ishida, Katsuya Takanashi, and Katherine Isbister. 2003. Can software agents influence human relations? Balance theory in agent-mediated communities. In *Proceedings of the second international joint conference on autonomous agents and multiagent systems*. 717–724.

[119] Jaya Narain, Tina Quach, Monique Davey, Hae Won Park, Cynthia Breazeal, and Rosalind Picard. 2020. Promoting wellbeing with Sunny, a chatbot that facilitates positive messages within social groups. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[120] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.

[121] Thien Hai Nguyen and Kiyoaki Shirai. 2015. Topic modeling based sentiment analysis on social media for stock market prediction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 1354–1364.

[122] Francisco Nunes, Nervo Verdezoto, Geraldine Fitzpatrick, Morten Kyng, Erik Grönvall, and Cristiano Storni. 2015. Self-care technologies in HCI: Trends, tensions, and opportunities. *ACM Transactions on Computer-Human Interaction (TOCHI)* 22, 6 (2015), 1–45.

[123] Magalie Ochs, Nathan Libermann, Axel Boidin, and Thierry Chaminade. 2017. Do you speak to a human or a virtual agent? automatic analysis of user's social cues during mediated communication. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 197–205.

[124] Shochi Otogi, Hung-Hsuan Huang, Ryo Hotta, and Kyoji Kawagoe. 2013. Finding the timings for a guide agent to interveneinter-user conversation in considering their gazebehaviors. In *Proceedings of the 6th workshop on Eye gaze in intelligent human machine interaction: gaze in multimodal interaction*. 19–24.

[125] Shochi Otogi, Hung-Hsuan Huang, Ryo Hotta, and Kyoji Kawagoe. 2014. Analysis of personality traits for intervention scene detection in multi-user conversation. In *Proceedings of the second international conference on Human-agent interaction*. 237–240.

[126] Leysia Palen and Paul Dourish. 2003. Unpacking" privacy" for a networked world. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 129–136.

[127] Endang Wahyu Pamungkas. 2019. Emotionally-aware chatbots: A survey. *arXiv preprint arXiv:1906.09774* (2019).

[128] Gilchan Park and Line Pouchard. 2019. Scientific Literature Mining for Experiment Information in Materials Design. In *2019 New York Scientific Data Summit (NYSDS)*. IEEE, 1–4.

[129] Gilchan Park, Julia Taylor Rayz, and Line Pouchard. 2020. Figure descriptive text extraction using ontological representation. In *The Thirty-Third International Flairs Conference*.

[130] Hyanghee Park and Joonhwan Lee. 2020. Can a Conversational Agent Lower Sexual Violence Victims' Burden of Self-Disclosure?. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.

[131] Florian Pecune, Jingya Chen, Yoichi Matsuyama, and Justine Cassell. 2018. Field trial analysis of socially aware robot assistant. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems*. 1241–1249.

[132] Zhenhui Peng, Taewook Kim, and Xiaojuan Ma. 2019. GremoBot: Exploring emotion regulation in group chat. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 335–340.

[133] Sandra Petronio. 2013. Brief status report on communication privacy management theory. *Journal of Family Communication* 13, 1 (2013), 6–14.

[134] Sandra Petronio and Wesley T Durham. 2014. Communication Privacy Management Theory. *Engaging Theories in Interpersonal Communication: Multiple Perspectives* (2014), 335.

[135] Laura R Pina, Carmen Gonzalez, Carolina Nieto, Wendy Roldan, Edgar Onofre, and Jason C Yip. 2018. How Latino children in the US engage in collaborative online information problem solving with their families. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–26.

[136] David Pinelle and Carl Gutwin. 2000. A review of groupware evaluations. In *Proceedings IEEE 9th International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET ICE 2000)*. IEEE, 86–91.

[137] Archana Prasad, Sean Blagsvedt, Tej Pochiraju, and Indrani Medhi Thies. 2019. Dara: A Chatbot to Help Indian Artists and Designers Discover International Opportunities. In *Proceedings of the 2019 on Creativity and Cognition*. 626–632.

[138] Paul Ralph and Romain Robbes. 2020. The ACM SIGSOFT Paper and Peer Review Quality Initiative: Status Report. *ACM SIGSOFT Software Engineering Notes* 45, 2 (2020), 17–18.

[139] Amon Rapp, Lorenzo Curti, and Arianna Boldi. 2021. The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots. *International Journal of Human-Computer Studies* (2021), 102630.

[140] Matthias Rehm. 2008. "She is just stupid"—Analyzing user–agent interactions in emotional game situations. *Interacting with Computers* 20, 3 (2008), 311–325.

[141] Samantha Reig, Michal Luria, Janet Z Wang, Danielle Oltman, Elizabeth Jeanne Carter, Aaron Steinfeld, Jodi Forlizzi, and John Zimmerman. 2020. Not Some Random Agent: Multi-person interaction with a personalizing service robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. 289–297.

[142] Minjin Rheu, Ji Youn Shin, Wei Peng, and Jina Huh-Yoo. 2021. Systematic review: Trust-building factors and implications for conversational agent design. *International Journal of Human–Computer Interaction* 37, 1 (2021), 81–96.

[143] Sherry Ruan, Jiayu He, Rui Ying, Jonathan Burkle, Dunia Hakim, Anna Wang, Yufeng Yin, Lily Zhou, Qianyao Xu, Abdallah AbuHashem, et al. 2020. Supporting children's math learning with feedback-augmented narrative technology. In *Proceedings of the Interaction Design and Children Conference*. 567–580.

[144] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L Murnane, Emma Brunskill, and James A Landay. 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.

[145] Benjamin S. Alpers, Kali Cornn, Lauren E. Feitzinger, Usman Khaliq, So Yeon Park, Bardia Beigi, Daniel Joseph Hills-Bunnell, Trevor Hyman, Kaustubh Deshpande, Rieko Yajima, et al. 2020. Capturing Passenger Experience in a Ride-Sharing Autonomous Vehicle: The Role of Digital Assistants in User Interface Design. In *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*. 83–93.

[146] Samiha Samrose, Kavya Anbarasu, Ajjen Joshi, and Taniya Mishra. 2020. Mitigating Boredom Using An Empathetic Conversational Agent. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[147] Shruti Sannon, Brett Stoll, Dominic DiFranzo, Malte Jung, and Natalya N Bazarova. 2018. How personification and interactivity influence stress-related disclosures to conversational agents. In *companion of the 2018 ACM conference on computer supported cooperative work and social computing*. 285–288.

[148] Kyle-Althea Santos, Ethel Ong, and Ron Resurreccion. 2020. Therapist vibe: children's expressions of their emotions through storytelling with a chatbot. In *Proceedings of the Interaction Design and Children Conference*. 483–494.

[149] Saiph Savage, Andres Monroy-Hernandez, and Tobias Höllerer. 2016. Botivist: Calling volunteers to action using online bots. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. 813–822.

[150] Daniel Schulman and Timothy Bickmore. 2009. Persuading users through counseling dialogue with a conversational agent. In *Proceedings of the 4th international conference on persuasive technology*. 1–8.

[151] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. 2018. " Hey Alexa, What's Up?" A Mixed-Methods Studies of In-Home Conversational Agent Usage. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 857–868.

[152] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–29.

[153] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond dyadic interactions: Considering chatbots as community members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[154] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. 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.

[155] Ameneh Shamekhi, Q Vera Liao, Dakuo Wang, Rachel KE Bellamy, and Thomas Erickson. 2018. Face Value? Exploring the effects of embodiment for a group facilitation agent. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[156] Weiyan Shi, Xuewei Wang, Yoo Jung Oh, Jingwen Zhang, Saurav Sahay, and Zhou Yu. 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.

[157] Hideo Shimazu. 2002. Expertclerk: a conversational case-based reasoning tool fordeveloping salesclerk agents in e-commerce webshops. *Artificial Intelligence Review* 18, 3 (2002), 223–244.

[158] Marita Skjuve, Ida Maria Haugstveit, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2019. HELP! Is my chatbot falling into the uncanny valley? An expirical study of user experience in human-chatbot interaction. *Human Technology* 15, 1 (2019).

[159] Thomas N Smyth, John Etherton, and Michael L Best. 2010. MOSES: Exploring new ground in media and post-conflict reconciliation. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 1059–1068.

[160] Constantine Stephanidis, Gavriel Salvendy, Margherita Antona, Jessie YC Chen, Jianming Dong, Vincent G Duffy, Xiaowen Fang, Cali Fidopiastis, Gino Fragomeni, Limin Paul Fu, et al. 2019. Seven HCI grand challenges. *International Journal of Human–Computer Interaction* 35, 14 (2019), 1229–1269.

[161] Sarah Strohkorb Sebo, Margaret Traeger, Malte Jung, and Brian Scassellati. 2018. The ripple effects of vulnerability: The effects of a robot's vulnerable behavior on trust in human-robot teams. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. 178–186.

[162] Lijun Sun and Yafeng Yin. 2017. Discovering themes and trends in transportation research using topic modeling. *Transportation Research Part C: Emerging Technologies* 77 (2017), 49–66.

[163] Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2015. Automated social skills trainer. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. 17–27.

[164] Stergios Tegos, Stavros Demetriadis, and Anastasios Karakostas. 2015. Promoting academically productive talk with conversational agent interventions in collaborative learning settings. *Computers & Education* 87 (2015), 309–325.

[165] Silke ter Stal, Lean Leonie Kramer, Monique Tabak, Harm op den Akker, and Hermie Hermens. 2020. Design features of embodied conversational agents in eHealth: a literature review. *International Journal of Human-Computer Studies* 138 (2020), 102409.

[166] Paul Thomas, Daniel McDuff, Mary Czerwinski, and Nick Craswell. 2020. Expressions of style in information seeking conversation with an agent. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1171–1180.

[167] Walter F Tichy, Paul Lukowicz, Lutz Prechelt, and Ernst A Heinz. 1995. Experimental evaluation in computer science: A quantitative study. *Journal of Systems and Software* 28, 1 (1995), 9–18.

[168] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. 2018. Understanding chatbot-mediated task management. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–6.

[169] Teun A Van Dijk. 1997. *Discourse as structure and process*. Vol. 1. Sage.

[170] Michelle ME Van Pinxteren, Mark Pluymaekers, and Jos GAM Lemmink. 2020. Human-like communication in conversational agents: a literature review and research agenda. *Journal of Service Management* (2020).

[171] Daniel Vogel and Ravin Balakrishnan. 2004. Interactive public ambient displays: transitioning from implicit to explicit, public to personal, interaction with multiple users. In *Proceedings of the 17th annual ACM symposium on User interface software and technology*. 137–146.

[172] Sarah Theres Völkel, Renate Haeuslschmid, Anna Werner, Heinrich Hussmann, and Andreas Butz. 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*. 1–15.

[173] Alexandra Vtyurina, Denis Savenkov, Eugene Agichtein, and Charles LA Clarke. 2017. Exploring conversational search with humans, assistants, and wizards. In *Proceedings of the 2017 chi conference extended abstracts on human factors in computing systems*. 2187–2193.

[174] Jacques Wainer and Claudia Barsottini. 2007. Empirical research in CSCW—a review of the ACM/CSCW conferences from 1998 to 2004. *Journal of the Brazilian Computer Society* 13, 3 (2007), 27–35.

[175] Jacques Wainer, Claudia G Novoa Barsottini, Danilo Lacerda, and Leandro Rodrigues Magalhães de Marco. 2009. Empirical evaluation in Computer Science research published by ACM. *Information and Software Technology* 51, 6 (2009), 1081–1085.

[176] Thiemo Wambsganss, Anne Höch, Naim Zierau, and Matthias Söllner. [n.d.]. Ethical Design of Conversational Agents: Towards Principles for a Value-Sensitive Design. ([n. d.]).

[177] Thiemo Wambsganss, Rainer Winkler, Matthias Söllner, and Jan Marco Leimeister. 2020. A conversational agent to improve response quality in course evaluations. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–9.

[178] Jinping Wang, Hyun Yang, Ruosi Shao, Saeed Abdullah, and S Shyam Sundar. 2020. Alexa as coach: Leveraging smart speakers to build social agents that reduce public speaking anxiety. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[179] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[180] Meredith Whittaker, Kate Crawford, Roel Dobbe, Genevieve Fried, Elizabeth Kaziunas, Varoon Mathur, Sarah Mysers West, Rashida Richardson, Jason Schultz, and Oscar Schwartz. 2018. *AI now report 2018*. AI Now Institute at New York University New York.

[181] Joost F Wolfswinkel, Elfi Furtmueller, and Celeste PM Wilderom. 2013. Using grounded theory as a method for rigorously reviewing literature. *European journal of information systems* 22, 1 (2013), 45–55.

[182] Ziang Xiao, Michelle X Zhou, Wenxi Chen, Huahai Yang, and Changyan Chi. 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.

[183] Ziang Xiao, Michelle X Zhou, and Wat-Tat Fu. 2019. Who should be my teammates: Using a conversational agent to understand individuals and help teaming. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 437–447.

[184] Ziang Xiao, Michelle X Zhou, Q Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell me about yourself: Using an AI-powered chatbot to conduct conversational surveys with open-ended questions. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 3 (2020), 1–37.

[185] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 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.

[186] Huichen Yang, Carlos A Aguirre, F Maria, Derek Christensen, Luis Bobadilla, Emily Davich, Jordan Roth, Lei Luo, Yihong Theis, Alice Lam, et al. 2019. Pipelines for procedural information extraction from scientific literature: towards recipes using machine learning and data science. In *2019 International conference on document analysis and recognition workshops (ICDARW)*, Vol. 2. IEEE, 41–46.

[187] Qian Yu, Tonya Nguyen, Soravis Prakkamakul, and Niloufar Salehi. 2019. " I Almost Fell in Love with a Machine" Speaking with Computers Affects Self-disclosure. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[188] Xiang Yuan and Yam San Chee. 2005. Design and evaluation of Elva: an embodied tour guide in an interactive virtual art gallery. *Computer animation and virtual worlds* 16, 2 (2005), 109–119.

[189] Amy X Zhang and Justin Cranshaw. 2018. Making sense of group chat through collaborative tagging and summarization. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–27.

[190] Jun Zheng, Xiang Yuan, and Yam San Chee. 2005. Designing multiparty interaction support in Elva, an embodied tour guide. In *Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*. 929–936.

[191] Qingxiao Zheng, Daniela M Markazi, Yiliu Tang, and Yun Huang. 2021. " PocketBot Is Like a Knock-On-the-Door!": Designing a Chatbot to Support Long-Distance Relationships. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[192] Michelle X Zhou, Gloria Mark, Jingyi Li, and Huahai Yang. 2019. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 9, 2-3 (2019), 1–36.

A APPENDIX

A.1 Definitions of Different Metrics Used in *Polyadic* Papers

Category	Metrics	Definition
Sociability	discussion quality	the quality of users' discussions
	consensus reaching	the reaching of a consensus in terms of behavioral and perceived opinion alignment
	opinion expression	users' message quantity, even participation, and perceived outspokenness
	opinion diversity	the number of unique lexical morphemes shared within a group
	even participation	how equally individual members contribute to the discussion.
	linguistic features	the linguistic features extracted from users' utterance text, e.g. verbosity, readability, sentiment, linguistic diversity and adaptability
Task-related	task completion time	the time elapsed before users successfully completed the intended task(s).
	efficiency	the number of system turns required and average system reaction time to users' requests.
	effectiveness	users' successes in the tasks undertaken with the CA support.
	satisfaction	users' overall ratings for CA's understandability, relevancy and efficiency
	team performance	the team's success in the tasks undertaken with the CA's support
Communication quality	perceived communication effectiveness	the degree of how much the CA helps users reach the goal
	perceived communication fairness	the degree of participation fairness that users perceive during discussions
	perceived communication efficiency	how quickly the consensus is reached during discussions
	perceived group climate	the relatively enduring tone and quality of group interaction experienced similarly by group members
Socio-emotional	perceptions of other group members	users' opinion and perception about other members in group discussions
	perception towards the agent	users' opinion and perceptions about the CA in discussion
	perceptions about the collaboration	users' collaboration experience in terms of users' awareness of each other's activities, effort, and enjoyment.
	perceived social presence	users' experience of being present with other persons and having access to their thoughts and emotions
	perceived social support	users' experience of being provided with support from other persons during discussion
	level of annoyance with the intervention	users' perceived annoyance when receiving the CA's intervention
	opinion of oneself/partner/cultural stereotypes	users' opinion about self, partner, and cultural stereotypes under the CA's influence during discussion
	perceived appropriateness of CA's social behavior	users' perception about the CA's ability to behave appropriately like a human
	psychological wellbeing	users' psychological wellbeing measured by positive relationships with others, personal mastery, autonomy, a feeling of purpose and meaning in life, and personal growth and development
Traditional	anthropomorphism	the extent to which the CA can demonstrate attribution of human characteristics or behavior
	perceived emotional competence	users' perception of the CA's emotional skills
	unmet expectations	users' expectations that are not met by the CA during the study
	privacy concerns	users' concerns about the safeguarding and usage of personal data provided to the CA
	level of perceived conversation control	users' feeling in control of the discussion using the CA
	perceived satisfaction	users' overall ratings for CA's understandability, relevancy and efficiency
	performance/effectiveness	the measure of user's success in the tasks undertaken in the CA interactions
	level of engagement	users' level of engagement with the CA, e.g., the number of interactions measured by clicks and selections
Sociability	perceived capability to promote contributions	users' perception about the CA's ability to elicit contributions and opinions from participants during discussion
Issue-relevant	decisions to (not) engage with the CA	users' decision of whether users would like to engage in interactions with the CA
	overall UX	users' perception of the overall user experience during interactions with the CA
	reflections on selves and others	users' reflections on behaviors and feelings of self or other participants
	willingness to take actions	users' willingness to take certain actions under the persuasion of the CA
	direction of improvement	the possible directions of improvements proposed by users' after interacting with the CA

A.2 Definitions of Different Metrics Used in *Dyadic* Papers

Category	Metrics	Definition
Linguistic features	positive/negative words	number of users' use of positive/negative words to express emotions
	length of utterances	users' number of users' use of words per statement
	personal pronouns	users' use of first- and third-person pronouns
	term-level repetition	the proportion of terms in one utterance which were repeated from the participant's previous utterance
	utterance-level repetition	the proportion of utterances where term-level repetition is greater than zero
	use of concrete words	the concreteness of each entry, e.g., "tennis" is more concrete than "sports"
	variation in language used	the variation in expressions and use of different words
Prosodic features	variation in speech-based agents	the word count, variance in pitch, rate and loudness in audio interactions
Dialogue features	dialogue/response quality	the syntactic readability and intensity of sentiments in users' replies
	dialogue efficiency metrics	the number of system turns required and average system reaction time to users' requests
	expressiveness	the quality of effectively conveying thoughts or feelings by users
Tasks related	task fulfillment rate/effectiveness	the measure of users' success rate in tasks undertaken during the interactions
	system use	a mixed measure of multiple system-related metrics, e.g., types of messages, words, time taken to complete the task
	user entry time	the amount of time users need to provide inputs to or interact verbally with the CA
	dropout rate	the percentage of respondents who quit before the study was completed
Algorithm	intent modeling evaluation	the accuracy of which the CA can identify the correct intents from users' utterances
	error rate	the rate of errors occurred during users' interaction with the CA
User	self-disclosure	users' quality of responses to the CA based on their trust towards the CA
	intimacy	users' perceive closeness,inter-connectedness, and companionship from the CA
	reflection	users' self-rated frequency of reflecting on thoughts and consciousness of their inner feelings
	impolite/aggressive behaviors	the number of occurrences of impolite phrases
	motivation	users' motivations or intents that drive users to interact with the CA
	affective states	the proportion of time users spent in each affective state

Category	Metrics	Definition
Conversation related	syntactic readability	the syntax-wise readability of conversation text between users and CA based on the Flesch-readability score
	self-reported response quality	the perceived response quality reported by users
	informativeness/information quality	the quality of information conveyed by the CA through text
	conversation smoothness	the level of smoothness of the conversation between users and CAs measure by Session Evaluation Questionnaire
	intensity of sentiments	the intensity of sentiments expressed by users calculated by TextBlob
	instruction quality	the quality of instructions given by the CA to users
Perception towards the CA	perceived common ground	the perceived mutual understanding between users and the CA
	opinion of the CA as a partner	the ease of which subjects were able to interact with the CA
	playfulness/enjoyment	the extent to which the CA can match users' interests
	sociability	users' perception of the CA's social skills
	perceived anthropomorphism	the extent to which the CA can demonstrate attribution of human characteristics or behaviors
	perceived warmth of the AI system	the extent to which users feel the CA is good-natured and warm
	impression	users' feelings of the CA regarding its competence, confidence, warmth, and sincerity
	perceived personality traits	users' feelings of the CA in terms of openness, conscientiousness, extraversion, agreeableness, and neuroticism
	perceived persuasiveness	users' perception of the CA's utterances rated as bad–good, foolish–wise, negative–positive, beneficial–harmful, effective–ineffective, and convincing–unconvincing
	perceived intimacy	users' perceived intimacy and closeness with the CA, e.g., feeling of interconnectedness of self and other
	perceived naturalness	users' agreement level to the statement that the CA's responses are natural
	perceived safeness	users' sense of safety when interacting with the CA
System usability	overall usability	users' perceived overall usability of the CA
	perceived ease of use	the degree to which users believe that interacting with the CA would be free of effort
	mental demand	the amount of mental or perceptual activities (e.g., thinking, deciding, calculating, remembering, looking, searching) that is required to interact with the CA
	physical demand	the amount of physical activities (e.g., pushing, pulling, controlling, activating, etc.) that is required to interact with the CA
	perceived task completion	the degree to which users believe to have successfully communicated and reached a mutual understanding with the CA
	effort	the total workload associated with the tasks, considering all sources and components
	frustration	users' feelings of being insecure, discouraged, irritated, and annoyed versus being secure, gratified, content and complacent when interacting with the CA
	desire to continue the interaction	the degree to which users would consider keep using this method in the future, or users' behavioural tendencies through their desire to help and cooperate with the CA
	user acceptance of the agent	users' willingness to accept the CA's interaction
	system consistency	the consistency between the behaviors and utterances of the CA
	perceived usefulness / helpfulness	users' own perceptions of the session's efficacy, e.g., the CA gave users good suggestions for helping them discover songs.
	willingness to recommend	the degree to which users would recommend the CA to their friends or family for managing mental well-being and to people who have needs
	motivation	users' motivation or intent to interact with the CA

Category	Metrics	Definition
User experience with the CA	user experience (UX)	a mix ratings of the general experience, e.g., emotion, ease of use, usefulness, and intention to use
	contrast of user experience	users' perceptions of the experience without drawing explicit attention to the contrast between their expectations and their experience
	perceived quality of the interaction	users' overall self-rated quality of the CA, e.g., in communicating, building rapport, and task fulfillment's
	perceived satisfaction	users' overall ratings for CA's understandability, relevancy and efficiency
	perceived engagement	the degree that the CA can engage the participants during the conversation, e.g., making users feel it is entertaining and interesting to engage in a dialogue with the CA
	perceived trust	the CA's ability in providing unbiased and accurate suggestions and making users trust it
	confidence	users' confidence that users will like the content the CA suggests
	perceived capability to control	users' feeling of being in control of the conversation
	aroused emotions	change of users' emotions state while using the system
	serendipity	the CA's ability in recommending things that users had not considered in the first place but turned out to be a positive and surprising discovery
	pleasant surprise	the CA's ability in providing contents that are overall pleasantly surprising to users
	empathy	the CA's ability to understand and share the feelings of users
	self-reflection/self-awareness	users' reflection on thoughts and consciousness of their inner feelings
	self-disclose	the degree of which users feel comfortable about disclosing to the agent and express opinions openly
	anxiety/tension	the degree of which the interaction makes users feel anxious or tense
	comfort	the degree of which the interaction is comfortable
	social orientation toward CAs	the desire to engage in human-like social interactions with CA, which is associated with a mental model of an agent system as being a sociable entity
	degree of collaboration	the level of collaboration between users and the CA during interaction
	degree of decision-making	the degree of which the CA helps with users' decision-making
Perceptions	overall impressions and experience	users' perceptions of the overall experience interacting with the CA
	perceived CA's capabilities	the CA's capability in handling simple requests and resembling human representative
	perceived CA appearance and self-presentation	the CA's visual appearance and persona
	perceived personality	users' feeling for the CA in openness (intellectual curiosity, creativity), conscientiousness (neatness, perseverance, reliability, and responsibility), extraversion (sociability, activity, and assertiveness), agreeableness (friendliness, helpfulness, and cooperativeness in dealing with others) and neuroticism (stability, anxiety, and the frequency of experiencing negative affect)
	perceived naturalness	the degree of which users feel the conversation with the CA is natural, not forced
	reciprocative	the degree of which users feel the CA reciprocated their language or feelings
	perceived burden	the degree of which users feel the conversation with the CA is costly in time, financially, mentally and emotionally.
	perceived human-likeness	the CA's ability to talk like a human and its conversational skills
	perceived effectiveness	the degree of how much the CA helps to address users' needs
Behaviors	engagement	the degree that the CA can engage the participants during the conversation, e.g., measured by an engagement rating between 1 (not engaging) and 5 (very engaging) in user survey
	notice of CA's certain function	the action of the CA sending messages informing what the chatbot could do, noticing a tutorial and menu
	efforts in learning the system	efforts users take in learning how to interact with the CA
	self-reflection/self-awareness	users' reflections on thoughts and consciousness of their inner feelings
	effects on judgement	the degree that users feel the CA affected their evaluation positively, negatively or neither
	effects on collaborations	the degree of which the CA affected users' willingness of collaborating with the CA
	daily practices of using the CA	participants' daily practices of using the chatbot, e.g., "Please briefly tell us how you used this chatbot during the past three weeks"