Chapter 7

Conversational Agents

*Ana Paula Chaves (Northern Arizona University, 1259 S Knoles Drive, Flagstaff, AZ, USA, Ana.Chaves@nau.edu) [0000-0002-2307-3099]

Charlotte van Hooijdonk (Utrecht University, Heidelberglaan 8, 3584 CS Utrecht, the Netherlands, C.M.J.vanhooijdonk@uu.nl) [0000-0001-5931-8417]

Christine Liebrecht (Tilburg University, Warandelaan 2, 5037 AB Tilburg, the Netherlands, C.C.Liebrecht@tilburguniversity.edu) [0000-0002-6621-2212]

Guilherme Corredato Guerino (State University of Maringá, gcguerino@uem.br) [0000-0002-4979-5831]

Heloisa Candello (IBM Research, hcandello@br.ibm.com) [0000-0002-4979-5831]

Minha Lee (Eindhoven University of Technology, the Netherlands, M.Lee@tue.nl) [0000-0002-7990-9035]

Matthias Kraus (Augsburg University, matthias.kraus@uni-a.de) [0000-0001-7400-2584]

Marco Aurelio Gerosa (Northern Arizona University, 1259 S Knoles Drive, Flagstaff, AZ, USA, Marco.Gerosa@nau.edu) [0000-0003-1399-7535]

Abstract: This chapter discusses how the ability to use (human) natural language has changed the way people interact with software systems. Navigating from Turing's Imitation Game to modern intelligent assistants, this chapter provides an overview of conversational agents in both technical and social dimensions of Human-Computer Interaction (HCI). It also explores the techniques and methodologies to develop and evaluate conversational agents under the lens of HCI and discusses the implications and future of these agents in human society.

1 Introduction

The American movie *her* (Jonze, 2014), directed in 2013 by Spike Jonze, displayed the dilemma of a professional writer who develops a romantic relationship with an AI assistant, Samantha. In the plot, Samantha is a powerful AI that can read books and review Theodore's letters in an incredibly short time but also portrays many human capabilities, such as using natural language communication, displaying emotions, and having an identity.

Her was not the first time the big screens raised the idea of AI agents co-existing with humans. Sci-fi movies and literature are full of examples of software that has a body and a soul, integrating into human society peacefully or disrupting this society by threatening human existence. Either way, when we reflect on these fictional scenarios, and the role AI currently plays in our day-to-day life, we may

wonder: are we walking into our most favorite (or feared) sci-fi plot? And, perhaps more importantly, how can we shape that future to achieve the most optimistic outcome?

The answer to these questions is not simple. When *Her* was released in December 2013, the virtual assistant Siri was already a commercial product built into Apple's product. Apple introduced the first iPhone featuring Siri in 2011, the same year that Watson, IBM's question & answer computer, won *Jeopardy! Challenge,* is a popular quiz television show in the USA. These events were not coincidental; they resulted from decades of research and industry initiatives to design human-like technology.

These initiatives were not sci-fi inspired, though. In 1950, Alan Turing's research manuscript "Can machines think?" argued that a machine can exhibit intelligent behavior indistinguishable from a human (Turing, 1950). This article would be the seed for the many research developments on artificial intelligence we have seen in modern computing. The initially called "imitation game" challenged scholars worldwide to create machines that could successfully play the game. To achieve that, machines would have to master one of the most defining features of humanity: natural language.

Significant efforts in natural language processing led to the scientific and technological advances that culminated in Siri and IBM Watson's spotlight in 2011. In the subsequent years, we watched the skyrocketing growth of conversational technologies, which would include intelligent personal assistants, home devices, virtual companions, and chatbots. Conversational products extrapolated the boundaries of specialized groups of users and research labs to reach the general public. Home personal assistants, such as Alexa and Google Home, reached people's homes and integrated people's daily routines (Dale, 2020). The accelerated digital transformation motivated by the COVID-19 pandemic boosted the spread of conversational agents to automate communication with customers in many service-based companies (Stoilova, 2021), such as customer service, healthcare, and financial services. ChatGPT, which is a conversational agent driven by Artificial Intelligence (AI), helps users to compose emails and essays, and even write code in particular programming languages, all based on natural language conversations (van Dis et al., 2023). The Conversational AI market has grown by billions of dollars (in USD), and the trend is estimated to continue in the upcoming years (Rashita et al., 2021).

Although talking to machines in natural language has become old news, the current scenario looks different from a sci-fi plot. Compared to Samantha, currently available conversational agents lack the complexity presented in Samantha's functionalities and personality. For the most part, conversational agents have been designed to accomplish narrow goals, for example, providing customer support for a particular company or assisting elderly individuals in their living environments. Even personal assistants, such as Siri and Alexa, who have a broader functional scope, do not replicate human communication in all its intricacy. As these technologies evolve to include more capabilities, scientists

and industry professionals take on the responsibility to create conversational agents that integrate with human activities in ways that promote a positive experience by increasing or controlling the extension of the conversational agent's humanity.

The Human-Computer Interaction (HCI) field plays a crucial role in the conversational agent's evolution. On the one hand, professionals in the area are concerned with the impact of functional performance on human perceptions, reactions, and inclination to adopt conversational technologies. On the other hand, as conversational agents present increased human features than traditional interfaces–at least the natural language capability, it is necessary to understand and shape the social and emotional expectations projected in the interactions with conversational interfaces.

Therefore, this chapter examines the role of HCI in the design and evolution of conversational agents. The chapter discusses the most significant aspects of human-conversational agents interaction, which include the process for conversation design, the technical and social factors that play a role in conversational interactions with computers, and techniques for evaluating conversational interfaces. In Section 7.2, the focus is the varying definitions of conversational agents, their characteristics, and their history. Section 7.3 discusses techniques for conversation design including stakeholder discovering, interface, and interaction design. Sections 7.4 and 7.5 approach the technical and social dimensions of conversational agent interaction. The former reviews the architectural aspects of conversational agents and introduces algorithms and methods to develop these agents. The latter switches gears toward the Computer As Social Actors paradigm and the user expectations of a conversational agent's social representation. Section 7.6 reviews the evaluation methods used to assess the pragmatic and hedonic qualities of conversational agents. Finally, Section 7.7 discusses the societal benefits and challenges of conversational agent adoption. Section 7.8 brings closing remarks and future directions.

2 Communicating with computers

Since Alan Turing proposed the Imitation Game (a.k.a. Turing Test), making computers that interact with humans through conversation has been a challenge for researchers (Turing, 1950). The first software to play the Imitation Game was ELIZA (Weizenbaum, 1966), followed by a series of early technologies such as TinyMud (Mauldin, 1994), SHRDLU (Winograd, 1971), and A.L.I.C.E. (Wallace, 2009). In 1991, Dr. Hugh Loebner started the first Loebner Prize Competition where every year the most human-like computer is rewarded.

In the early 2000s, the popularity of conversational agents increased with their integration into instant messaging tools (Adamopoulou & Moussiades, 2020). Differently from the original agents, whose main goal was to mimic human conversations, conversational agents integrated with instant

messaging tools helped people with practical daily tasks such as retrieving information from databases about movie times, sports scores, stock prices, news, weather, etc. This ability marked a significant development in human-computer interaction as information systems became accessible through dialog.

The development of conversational agents went one step further with the creation of smart personal voice assistants in the early 2010s. These agents were built into smartphones or dedicated home speakers and could understand voice commands, speak to the user by synthesizing a voice, and handle tasks such as setting up alarms, monitoring devices, accessing calendars, etc. The most popular voice assistants are Apple Siri, IBM Watson, Google Assistant, Microsoft Cortana, and Amazon Alexa (Adamopoulou & Moussiades, 2020; Dale, 2020). In that same decade, social media platforms allowed companies to create chatbots to represent their brand or interact with their customers via a conversational interface. For example, airlines started to answer complaints on social media with chatbots and offer services such as checking in, informing about flight delays, and providing boarding passes (Ukpabi et al., 2019). Conversational agents automated a large number of messages that were previously answered by humans and thousands of text-based chatbots were developed for popular messaging platforms.

In the 2020s, advances in machine learning allowed the training of large-scale language models, which considerably improved the performance of conversational agents and allowed the integration of the technologies in a variety of applications. ChatGPT, one of these new conversational agents, has become notorious for its disruptive performance in a variety of contexts, including text generation, programming, and language translation (van Dis et al., 2023). The model is trained on a massive amount of data, allowing it to generate text that is often difficult to distinguish from text written by a human.

With the increasing popularity of human-computer interaction via conversational interfaces, a number of terms have been coined to accommodate the varying characteristics of these agents. In the following, we present an overview of these terminologies and discuss the domains where conversational agents have been mostly applied.

2.1 Definitions and classification

Generally speaking, conversational agents are “computer programs that interact with users using natural language”. Although the concept emerged several decades ago and has evolved over the years, the "conversational agent" term is not consensual (Motger et al., 2022) and it has been either used as a generic term for a software-based dialog system or as a synonym of chatbot. Several synonyms emerged, such as social bots, personal assistants, and conversational interfaces with the goal of emphasizing specific characteristics of the agents. In this chapter, conversational agents

represent a software application that uses natural language as the main form of interaction with humans.

Conversational agents can serve various purposes, with characteristics that distinguish one agent from the others, including physicality (dis/embodiment, animation, or avatars), input mode (voice- or text-based), goal (general-purpose chat or task-oriented), and domain. In general, conversational agents may interact through typed text, speech, or both means (Diederich et al., 2019). The interaction may be controlled by the user (user-initiative), by the conversational agent (system-initiative), or by both (mixed-initiative) (Motger et al., 2022). Conversational agents may also have personality, humor, and be able to express emotions; sometimes they have a face, or even a body (Feine et al., 2019). When conversational agents have physical characteristics, they are usually called "embodied conversational agents'' (Cassell, 2000). According to Cassell (2000), embodied conversational agents have the same properties as humans in face-to-face conversations, for example, they are able to recognize verbal and non-verbal inputs, answer with verbal and non-verbal outputs, and understand turn-taking. This chapter does not discuss embodiment (e.g., virtual humans or robots), although disembodied conversational agents may also have such an appearance (e.g., an avatar) (Feine et al., 2019).

2.2 Application domains

Conversational agents have changed how companies engage with their customers (Ling et al., 2021), how students participate in their learning groups (Khosrawi-Rad et al., 2022), and how patients self-monitor the progress of their treatment (Milne-Ives et al., 2020), among many other applications. Recent reports on the conversational agent market (Rashita et al., 2021) attest to their increasing demand in several different domains. According to Rashita et al. (2021), the domains where conversational agents are more expressively adopted include Bank, Financial Services & Insurance (BFSI), retail & e-commerce, healthcare & life science, travel & hospitality, telecom, and media & entertainment, with retail & e-commerce being the most expressive segment and anticipated to witness significant growth in the upcoming years.

Aligned with the marketing trend, many studies in the literature focus on the support conversational agents can provide to customer services (Følstad & Skjuve, 2019; Haugeland et al., 2022; Youn & Jin, 2021) and marketing (Cui et al., 2017; Kaczorowska-Spychalska, 2019; Thomaz et al., 2020). In the BFSI domain, conversational agents support decision-making and investment choices as well as productivity (Sharma et al., 2021; Wube et al., 2022). When focusing on travel & hospitality, conversational agents have been widely used in several subdomains, such as planning, booking, and en-route experience (Mahmood et al., 2009; Pillai & Sivathanu, 2020; Yanishevskaya et al., 2019). In health, they are used to enable speech monitoring, identification of disease, diabetes monitoring/control, and personal healthcare assistance (Laumer et al., 2019; Milne-Ives et al., 2020;

Preum et al., 2021).

There are several other domains where conversational agents have been investigated. For example, in the educational context, conversational agents have been widely adopted, with applications including tutoring, question-answering, conversation practice for language learners, learning companions, and dialogues to promote reflection and metacognitive skills (Hwang & Chang, 2021; Khosrawi-Rad et al., 2022; Wollny et al., 2021). In Software Engineering, Storey and Zagalsky (2016) attest that conversational agents are present in every phase of the software development process, including coding, testing, documentation, deployment, support, and even team coordination. GitHub, the most popular code hosting platform, has been swarmed with bots and chatbots that help developers in their daily tasks, interacting with the team via communication channels like comments on pull requests (Wessel et al., 2018). Since the Internet of Things brings internet connection to a variety of different physical devices or things, conversational agents have been studied as a solution to facilitate management and interaction with those devices (Augustsson, 2019).

2.3 Conversational agents and the HCI

Conversational agents are changing the patterns of interactions between humans and computers. Lugger and Sellen (2016) claim that "conversation is the next natural form of Human-Computer Interaction." Many instant messenger tools and social networking platforms provide platforms to develop and deploy conversational agents, which organizations use to provide their services (Følstad & Brandtzæg, 2017). As messaging tools and social network sites increasingly become platforms, traditional websites, and apps are providing space for this new form of human-computer interaction. The increasing interest in conversational technologies has brought new challenges for the HCI field (Følstad & Brandtzæg, 2017; Neururer et al., 2018). Whereas traditional user interfaces apply visual elements such as buttons, menus, or hyperlinks to communicate with users, conversational interfaces rely almost entirely on language as the primary resource to achieve communicative goals. Nevertheless, Dale (2016) states that interacting with current conversational agents conveys the impression of *being managed through a tightly controlled dialog flow* with reduced interactivity, which turns users into option-selectors rather than conversational partners.

Moreover, language design for conversational agents has focused primarily on ensuring that the agents produce coherent and grammatically correct responses, and on improving functional performance and accuracy (see e.g. (Jiang & E Banchs, 2017; Maslowski et al., 2017). Although current conversational agents may, at some functional level, provide users with the answers they seek, the utterances portray arbitrary patterns of language that often fail to take into account the interactional situation in choosing a proper conversational tone for the interaction. For example, the literature shows that appropriate linguistic choices potentially increase human likeness (Chaves et al., 2022; Hill

et al., 2015) and believability (Westerman et al., 2019; Xuetao et al., 2009), enhancing the overall quality of the interaction (Chaves et al., 2022; Jakic et al., 2017). Hence, making a conversational agent acceptable to users is not only a technical but also a social problem to solve (Neururer et al., 2018). Developing a strong basis for designing not just what a conversational agent says but also how it says it must be a priority for creating the next generation of human-computer interfaces.

Conversational agents are typically designed to mimic the social roles usually associated with a human conversational partner. Research on mind perception theory (Heyselaar & Bosse, 2019; M. Lee, Lucas, et al., 2019) suggests that although artificial agents are presumed to have substandard intelligence, people still apply certain social stereotypes to them. It is reasonable, then, to assume that ``machines may be treated differently when attributed with higher-order minds'' (M. Lee, Lucas, et al., 2019). As conversational agents enrich their communication and social skills, user expectations will likely grow as the conversational competence and perceived social role of the agents approach the human profiles they aim to represent. A variety of factors influence how people perceive an agent's communication skills (Chaves & Gerosa, 2020; Feine et al., 2019) and, as user expectations of proficiency increase, one important way to enhance the agent interactions is by carefully planning the technical, social and interactional aspects of the conversational agent design. In the next sections, we navigate these dimensions to discuss techniques applied to improve conversational agents' design.

3 Conversation design

The design of conversational agents requires specific tools and methods to comply with the unique characteristics of these systems. In this section, we will discuss techniques to design conversational agents, which includes first a preliminary analysis covering contextual and stakeholder discovery, approaches, and techniques. Second, we will rely on an overview of dialogue design methods and techniques. Implementation and evaluation techniques are described in Sections 7.4 Building Conversational Agents and 7.6 Conversational Agents Evaluation, respectively.

3.1 Contextual and stakeholder-discovering approaches and techniques

The ecosystem of designing conversational AI agents includes several components and resources. Users, developers, content curators, project managers, designers, machine learning analysts, and marketing and consulting employees are usually involved in the process. The diversity of stakeholders brings a lot of complexity, and a successful conversational agent design depends upon the alignment of several points.

Many theories, frameworks, and approaches assist researchers to understand and identify the nature and challenges of designing conversational systems and to assist the development team in making ethical (Rakova et al., 2021), and explainable design decisions (Ehsan et al., 2022; Q. V. Liao et al.,

2020). The nuances of stakeholders' views are not always clear for practitioners, designers, and developers. Understanding stakeholders' goals and mental maps helps to build the process of designing effective CA systems. This section addresses two approaches towards considering diverse lenses in the CA design process: value-sensitive design and articulated work practices.

A Value-Sensitive Design (VSD) methodology consists of integrating conceptual, empirical, and technical investigations (Friedman & Hendry, 2019). In this approach, value is defined as "what a person or group of people consider important in life" (Friedman et al., 2013). In this frame, the notion of direct and indirect stakeholders is considered. For example, users are generally considered direct stakeholders and project sponsors could be considered indirect stakeholders. Or even, developers sometimes could be the direct stakeholders when they are considered the users of conversational design platforms, and indirect stakeholders when they are developing the conversational system for end users. In all those situations, the values of stakeholders count to have a successful conversational system.

Wambsganss (2021) used a value-sensitive design approach (Friedman & Hendry, 2019), and design science research (Hevner, 2007) to investigate conversational agents' design principles. The authors found that this approach is suitable for user scenarios where privacy and transparency play an important role. Their work can inform designers and stakeholders about an ethical way to design conversational agents.

Görnemann & Spiekermann (2022) used the VSD approach along with other fields to create a framework called EVA (Emotion Value Assessment) aimed to assist researchers and practitioners in the HCI domain unveil the emotional reactions of voice-based conversational agents in relation to underlying values fostered or harmed. Understanding and grasping values and emotional reactions to technology can assist in developing ethical CA.

Values such as safety are also a great concern for CA researchers in end-to-end (E2E) conversational AI systems. Dinan et al. (2021) and Bergman et al. (2022) investigate how and when conversational agents trained on large datasets from the internet should be released considering safety, and value tensions in training and releasing E2E CA models. The authors discuss the uncertain nature of large language models operating under conversational user interfaces and categories of harmful responses. Those issues require weighing conflicting, uncertain, and changing values.

Values embedded in the design process interfere with decision-making tasks that are not always discussed and transparent for the CA team and users. For example, data curation and machine learning decisions (Rattenbury et al., 2017; Seidelin et al., 2020) are paramount steps of the CA design process and directly affect the quality and user perception of a CA. Usually, this work is hidden from stakeholders and is invisible to users (Ju & Leifer, 2008). We could say that those hidden practices are

a form of “articulated work”, a "work that is necessary for the work to proceed" (Hampson & Junor, 2005; Schmidt & Schmidt, 2011). Articulation work is usually a complex set of enabling activities that contribute indirectly to the more visible and prominent production work in workplaces (Candello et al., 2022). It can take the form of support work or simply invisible work that usually goes unobserved and unrecorded. Therefore, it is important to make this work visible (Nardi & ENgeström, 1999) and understand the nature of technology, and sometimes limited decisions included in the process. Several studies are uncovering the production work entitled developing conversational agents.

Candello (2022) investigated articulated working practices of content curators of conversational agents in diverse industry settings. They proposed a distinction between tech curators, the ones with knowledge of conversational platforms, and content curators, the subject matter experts, usually responsible for the content. In their research, three main themes emerged that illustrate the working practices of those employees: (1) co-dependence work mechanisms of curators

(2) cooperation and collaboration mechanisms, and (3) management of information spaces and technologies. From the study results, they also draw seven design implications to improve the interface design and features of conversational platforms.

In the same line of studies to unpack practitioners' work for conversational agents, Khemani and Reeves (2022) interviewed nine Voice-User Interface (VUI) practitioners to understand how they conceptualize and use design guidelines developed in HCI. One of the main takeaways from this study is that design knowledge should be codified for practitioners to be adopted. Lack of adoption was also found in a study based on a large-scale survey with 105 industry designers (Murad et al., 2022). The study aimed to explore the design practices of VUI designers, and in which ways knowledge of GUI is applied to designing VUI.

3.2 Dialogue design methods and techniques

A conversation is a specialized form of interaction, according to Suchman (2007, p. 101) *“a distinguishing feature of ordinary conversation is the local, moment-by-moment management of the distribution of turns, of their size, and what gets done in them, those things being accomplished in the course of each current speaker’s turn.”* Management of turns e subject change in each course is a situation that occurs in real-life conversations based on circumstances (internal and external) to speakers in dialogue. Machines are not prepared, nowadays, to fully understand the context and change the course of conversations as humans. Managing dialogues with machines is challenging, and this challenge increases even more when more than one conversational agent is part of the same conversation.

In a conversational user interface, the conversation flow is not linear or transparent; it might take different courses according to circumstances that influence dialogue. One of the most interesting

features of human conversation is the ability to explore sidetracks and easily go back to the main conversation objective. For instance, while people are making a decision, such as in planning a trip, people can ask clarifying questions, explore a similar case, get delighted by photos and comments, consult a friend, and then go back to make the trip decisions. It is almost impossible to predict in what sequence a user will interact with a machine and how this machine could provide a satisfactory user experience. Traditional design methods might help to envision graphical user interfaces and detect topics that will embody the conversational system but have clear limitations in supporting the design of the conversation flow of a conversational system. This section explores some design methods which may help in this process.

Zue and Glass (2000) addressed some of the challenges in designing dialogue flow. According to them, system-initiative systems (see Section 7.2.1 Classification) restrict user options, asking direct questions, such as: “Please, say just the departure city.” By doing so, those types of systems are more successful and easy to answer as they guide the user through the expected path. On the other hand, user-initiative systems are the ones where users have the freedom to ask what they wish. In this context, users may feel uncertain of the capabilities of the system and request information or services which might be quite far from the system's domain, leading to user frustration. In the mixed-initiative approach, in which users and computers participate interactively to achieve a common goal, the challenges include understanding interruptions, human utterances, and unclear sentences that were not always goal-oriented.

The key dilemma is: should we ask users to modify their behaviors and interact with the system in a structured way? Or should we let users be more comfortable with systems that have characteristics of humans? Or both? In our experience, this is in fact one of the earliest and often the most important decisions faced by designers of conversational systems and should be explored in the design process by, for instance, using a Wizard of Oz approach (Mateas, 1999) and prototype techniques.

In WoZ experiments, users are told they are interacting with a conversational agent, though in fact, a human plays the role of the agent behind the scenes. Samsom and Sumi (2020) conducted a WoZ experiment to understand the driver's decision-making process when listening to route alternatives in a conversation between two conversational agents. Chaves and Gerosa (2018) used a WoZ setup to identify differences in turn-taking and dialogue flow structure when users interact with single or multiple chatbots in the same chat.

Cambre and Kulkarni (2020) mentioned additional prototyping techniques and highlighted that after the elicitation phase, the emphasis may shift to a more exploratory design stage, where the prototypes resemble the final voice artifacts intended to be developed. In this case, conversational development platforms are used. Shorter et al. (2022) investigate *prototyping* as a design tool for materializing voice

assistant technologies, those approaches show the peculiarities of designing for intangible technologies. Following the approaches to use prototypes as prompts for designing voice assistants in-car settings, Meck (2022) conducted an experiment to compare three evaluation conditions: driving simulator; crowdsourcing audio, and crowdsourcing text. They discovered study participants processed prompts similarly in these conditions.

With the increased amount of information nowadays, designers of conversational agents struggle to organize the content in a user-friendly way. Several methods borrowed from GUI context can be adapted to the CA context. Methods, such as *card sorting* (Nawaz, 2012) help to organize and evaluate the information of an interface. For instance, in a call-center context, where several topics could be asked to the CA, designers can prioritize the topics by importance using a card sorting method with branch employees and call-center employees to identify the popular topics requested by customers.
Menus are also applied to show users the content scope of the CAs, and to organize the information. Nguyen et al. (2021) conducted an empirical study and found that CAs lead to a lower level of perceived autonomy and higher cognitive load compared to menu-based interface systems, resulting in lower user satisfaction. Considering Nguyen et al.'s (2021) findings, it is comprehensible that many designers are using those GUI strategies in a conversation interface. Hu (2019)compared the use of a menu-based over a conversational chatbot experience. In the study, users preferred the menu-based experience for several reasons such as being easier to use, less likelihood of errors, the convenience of GUI elements, and a clear way to show where information needs to be provided rather than requested. Valério (2020) investigated how communicative strategies are used by popular chatbots when conveying their features to users. They identified that menus offer the possibility of quick replies and choices of predefined options, complementary to Hu's (2019) study results. Menus also can help designers and developers to scope the system knowledge and keep the subject of the conversation in the scope, avoiding conversation breakdowns due to the lack of CA understanding.
Scenarios and Storyboards (Carrol, 1999; Llitjós, 2013) based on *journey maps* (Schneider & Stickdorn, 2011) and *blueprints* (Polaine et al., 2013) assist in predicting the user experience. Design fiction studies (Blythe, 2014) are also applied in designing future experiences with CA. Muller and Liao (2017) offered four potential methods to envision the values and ethical implications of designing AI systems for future users. Research to mitigate harm included fictional scenarios to assist practitioners to reflect when preparing for and learning from AI conversational models release (Bergman et al., 2022). More specifically to digital personal assistants, Søndergaard and Hansen (2018) discuss the meanings of designing and adopting CAs using design fiction through a critical and feminist design methodology. Ringfort-Felner et al. (2022) brought a collaborative and social perspective of using a design fiction artifact named Kiro in a car context. Participants in the study accepted Kiro as a conversational partner but not as a replacement for a human.

Designers should also consider examining similar contexts where everyday activities happen without or with the use of CAs to understand the nuances and values in place. Observations and log analysis of call-center employees, for example, may assist designers to understand the dynamic of the conversations, and design the dialogue for user experience. Porcheron et al. (2018) collected and analyzed audio data of VUI in participants' homes to understand the social interaction implications in everyday life. Barth et al. (2020) conducted a log analysis collected from visitors from an art exhibition to identify categories of popular questions. Portela and Granell-Canut (2017) conducted observations of chatbot mobile phone users to understand the potential personal relationships users might develop interacting with chatbots. Those studies are examples of how to conduct observational studies to gather insights for designing the interaction and conversational flow of CAs.

Another component of this decision is the technology available for the deployment of the platform. Different conversational platforms support different initiative models, so the designer may face application contexts where the initial strategy is predetermined by the platform. In this situation, they should focus on finding and identifying patterns of dialogue that make sense given a fixed initiative strategy. For instance, if the only available platform has a Q&A structure (a typical user initiative), the designer should consider answers which lead to specific questions from the users if more guidance to the user is needed. In any case, since the decision of the initial strategy is closely tied to the deployment platform capabilities, it is important to involve the developing team in the process and, often, make that decision as early as possible in the design process. The next section will cover the technology and techniques available to create conversational agents.

4 Building Conversational Agents

For being able to have a conversation with a computer system, several techniques from natural language processing (NLP), a subfield of AI, are combined to understand and respond appropriately to user input. These include natural language understanding (NLU) for determining the semantic meaning or intention of the user's utterance, dialog management (DM) for keeping track of conversational context and deciding on system actions, and natural language generation (NLG) for transforming abstract representations of system actions into natural language.

Typically, these techniques are implemented into a dialog system forming the technical foundation of conversational agents. For realizing dialog systems there exist different reference architectures that have been historically categorized according to their purpose. A modular architecture is usually applied to implement task-oriented dialog systems which are designed to assist users in achieving a predefined task or goal. For example, task-oriented systems have been used to realize conversational agents to handle hotel room bookings or for ordering food from a restaurant (Williams et al., 2016). A

characteristic of task-oriented oriented systems is that modules for NLU, DM, and NLG are implemented and fine-tuned separately. The purpose of NLU is for the system to understand a narrow range of user intents which are relevant for solving the task at hand. DM is realized using a rule-based, state machine, or statistical approaches for guiding the user to task accomplishment by following a predefined set of actions. Moreover, the DM module interacts with structured data sources, such as databases or APIs, to provide information and complete tasks. NLG is then used to present the actions in a human-understandable format. A benefit of this approach is that already implemented modules can be reused fast and easily. Further, the integration of external services for handling individual NLP tasks is facilitated and the architecture is easily extensible. This makes the modular approach especially useful for rapid prototyping aiming at investigating novel interaction design techniques or collecting data in new task domains. As the work processes of modular-based systems are easy to follow and technical problems can be alleviated relatively easily, this architectural design is popular for developing commercial voice user interface applications. For example, the modular-based RASA[1] framework can be used to build customer service chatbots, while Amazon also provides a modular framework for developing Alexa Skills.

Due to their ability to only solve narrow tasks and to understand a limited set of user intents, the modular architectural design is not suitable for handling open-domain or casual conversations with users on a wide range of topics. Therefore, End-to-End architectures are used to enable open-ended conversations between humans and artificial agents. This architectural design relies on neural dialog approaches for handling NLU, DM, and NLG, such as a sequence-to-sequence model like Long Short Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) or transformer-based model (Vaswani et al., 2017). These models are trained on large datasets of conversational data and learn to generate responses that are contextually relevant and engaging to users. For this, they either make use of retrieval-based methods, selecting existing utterances from a set of appropriate responses, or generative approaches, which model new utterances dependent on language models. Recently, End-to-End approaches have garnered widespread popularity for realizing chatbots able to handle chit-chat or question answering, e.g., Google's Meena (Adiwardana et al., 2020), Facebook's Blenderbot (Roller et al., 2021), or ChatGPT[2]. Despite their popularity, End-to-End systems still have several problematic issues (McTear, 2020). One of the main problems is the generic response problem, which concerns the often bland or uninformative responses of such systems, e.g., "Ok." or "I'm not sure.". Further, they are prone to semantic inconsistencies, i.e., their responses are inconsistent with their previous responses. For example, they may state different cities when asked about their current

[1] www.rasa.com

[2] https://openai.com/blog/chatgpt

habitat. The probably most severe problem are so-called hallucinations, where the system adds false information to their responses. For example, retrieval-based methods mirror responses from their training data and are thus prone to bias or false information existing in the data set. Similarly, generative models create new utterances based on knowledge about the properties of language rather than truly understanding the meaning of their responses and are thus prone to making up facts. However, due to their ability to generate human-like responses and being able to process a wide range of topics, open-domain systems can be quite entertaining and useful for handling constrained predictable interactions. Thus, there also exist approaches to make task-oriented systems more natural by combining modular and End-to-End models. For example, Bordes et al. (2017) studied the application of End-to-End models for task-oriented dialogue in the restaurant reservation domain. In the following, the two architectural designs are described in more detail.

4.1 Modular Architecture (600)

Task-oriented conversational agents are usually built using a modular architecture paradigm. A key characteristic of the modular architecture is the pipeline-like structure, where task-specific modules process user input and generate appropriate system responses. In Figure 7.1, a depiction of the typical architecture is presented. For voice-based systems, like Alexa, or Siri, the user's speech signal is first transformed into text using an automatic speech recognition (ASR) module. For this, the speech signal is first digitalized and pre-processed for removing noise and redundant information (Jurafsky & Martin, 2023). Afterwards, features according to the human auditory system are extracted and fed as input to a neural end-to-end model for generating the most probable word sequence based on the acoustic feature vector (Yu & Deng, 2016). Besides semantic and syntactic information, other relevant information for HCI can also be extracted from the speech signal, e.g., sentiments, or emotions, as well as the age, or gender of the user.

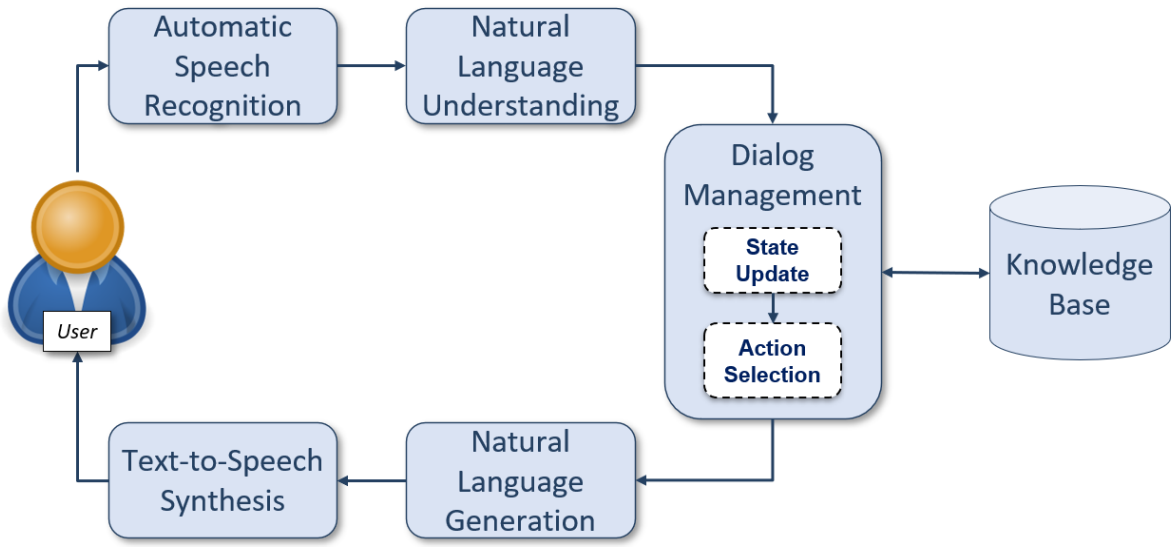

Figure 7.1. Modular dialog system architecture (McTear,2020).

Natural language understanding (NLU) describes the process of extracting the meaning of user input from a word sequence. For this, a semantic representation of the utterance is generated based on formal structures. This is also known under the term semantic encoding. Historically, semantic grammars have been used that structure the utterance according to the communicative function of their constituents (Traum & Hinkelman, 1992). For example, an utterance may have an assertive function, i.e., addressing the state of a current situation, or a directive function, i.e., intending to commit the addressee to do something (Searle, 1969). There exist several taxonomies of communicative functions that define specific dialogue actions (often also named intent) for semantically analyzing user and system turns during dialogue, e.g., DAMSL (dialog act markup in several layers) (Core & Allen, 1997). Therefore, the task of semantically encoding a dialogue utterance is described as dialogue act classification. Here, each utterance of a dialogue corpora is associated with a specific dialogue action (intent). Afterward the corpus is split into training, test, and validation set, for training a dialogue act classifier that maps the input word sequence to a semantic representation. Training is conducted by first generating a numerical representation of the utterance using a contextual embedding network, e.g., BERT (Devlin et al., 2019) or Word2vec (Mikolov et al., 2013), and then feeding the representation into a machine learning classifier, e.g., support vector machine (SVM) (Vapnik, 2000) or deep neural networks, for a 1-of-N classification of the dialogue act. In Table 7.1, a sample dialog from the HIS system by (Young et al., 2010) is presented which shows the mapping from utterances to dialog acts. Besides the dialogue act, also the task domain of the utterance can be determined in this way for systems that are able to handle multiple domains, e.g. see the first turn in Table 7.1, where the restaurant is determined as the domain. Further, relevant information on the word level can be classified, which is known under the term entity extraction. This is particularly useful for slot-filling dialogues to determine the values for specified slots. For this, BIO-tagging can be used that associates a tag for the beginning (B) and inside (I) for each slot label, and one for tokens outside (O) of any slot label (Jurafsky reference). For example, consider the second turn presented in Table 7.1, for which the BIO-tagging would be as shown in Figure 7.2:

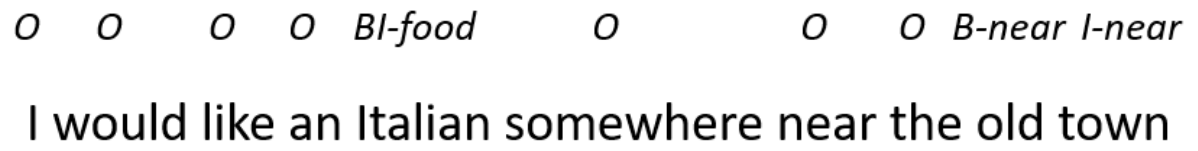

Figure 7.2. Example for BIO-Tagging. Created by the authors.

Table 7.1 Example dialog showing the mapping of utterances to dialog acts (Young et al., 2010)

Utterance	**Dialog act**
U: Hi, I am looking for somewhere to eat.	hello(task = find,type=restaurant)
S: You are looking for a restaurant. What type of food do you like?	confreq(type = restaurant, food)
U: I'd like an Italian somewhere near the old town.	inform(food = Italian, near=old town)
S: Napoli is a nice Italian restaurant near the old town.	inform(name = "Napoli", type = restaurant, food = Italian, near = old town)
U: Is it reasonably priced?	confirm(pricerange = moderate)
S: Yes, Napoli is in the moderate price range.	affirm(name = "Napoli", pricerange = moderate)
U: What is the phone number?	request(phone)
S: The number of Napoli is 764503.	inform(name = "Napoli", phone = " 764503")
U: Ok, thank you goodbye.	bye()

Dialogue management is concerned with finding an appropriate system response given the last user utterance and all available information from the previous interaction. Hence, its task is to keep track of the conversation and decide which actions to take for accomplishing a predefined task. To achieve this, a dialogue management component comprises a dialogue state tracker for maintaining the dialogue state, and a dialogue policy for controlling the flow of the dialogue (Young et al., 2013). The dialogue state tracker uses several knowledge sources for modeling the state of the conversation. The dialogue history contains information about the user's and system's contributions to the conversation thus far, i.e., which dialogue actions have been used and which values have been provided for slots. For example, food type, location, price range would be the slots which the system from Table 7.1 requires for providing adequate information. The domain model represents the "world knowledge" of

a system, i.e., concepts and information for a task domain, such as different types of food or locations in the restaurant search domain. This knowledge can be retrieved from a database, which can be structured, for example, in the form of a knowledge graph or an ontology. In addition to conversation- and domain-related knowledge, the dialog state can also contain user-specific information, e.g., age, gender, or preferences, as well as relevant dynamic user states, e.g., the classified sentiment of the user utterance. The dialogue policy determines what action the system should take next. For decision-making, the system makes uses of the relevant information represented in the dialogue state and selects an appropriate dialogue action. These decisions can be made using rule-based mechanisms or statistically driven methods. An example for a rule-based policy would be that when the confidence value for the user's providing the value to a specific slot is below a pre-specified threshold, the system asks for clarification. Otherwise, the system would proceed to ask for another slot or suggest, etc. In Table 7.1., for example, the system recognizes with a high probability that the user wants to go to a restaurant and therefore proceeds by asking which type of food the user wants to eat. If the confidence would have been low, the system could ask explicitly for confirmation, e.g. "You want to go to a restaurant. Is this correct?".

Statistically-driven systems automatically learn such decisions based on data corpora using supervised learning mechanisms or interactively using reinforcement learning. For example, Young et al. (2013) use an approach that learns a dialogue strategy automatically depending on the last user input and dialogue state represented as the slot-value pairs provided during the current conversation. For learning a strategy, the system receives a positive reward at the end of dialogue when all slots have been filled correctly by the system, and a negative reward otherwise. Further, a small negative reward can be given for each dialogue turn to learn efficient system behavior.

Natural language generation provides a word sequence given the semantic representation of the chosen system action. Thus, it can be considered as the reverse task to NLU. For achieving this, two approaches are primarily used: a template-based approach uses hand-crafted mappings of dialogue actions, slots, and values to textual utterances, which can be realized in the form of a look-up table. A more sophisticated approach to create more diversified text is to use statistical methods based on large hand-labeled dialogue data (Budzianowski et al., 2018). For this, several approaches have proven to provide good results including neural approaches (Dušek et al., 2020) as well as reinforcement learning approaches (Rieser et al., 2014).

If the system output is to be in natural language, a synthesis module uses the text representation to generate speech signals. Therefore, it is intended to solve the reverse task of ASR. Like ASR systems, text-to-speech synthesis (TTS) relies on neural End-to-End models, using LSTMs or Transformers.

However, the main difference between ASR and TTS concerns the training procedure. While the ASR needs to be trained speaker-independently for being able to recognize speech from various users, the TTS module is usually trained on a specific speaker for having a consistent voice (Jurafsky & Martin, 2023). For transforming text into speech, the process generally involves three tasks: First, text needs to be normalized for handling non-standard words, i.e., numbers, dates, abbreviations, etc. In a second step the sequence of normalized words is transformed into a numerical representation and fed as input to an encoder-decoder model that generates the predicted Mel spectrum of the spoken utterance (frequency spectrum) dependent on the input representation. Finally, vocoding is used to transform the spectral features back into the time-domain waveform representation which can be played back to the user.

4.2 End-to-End Architecture

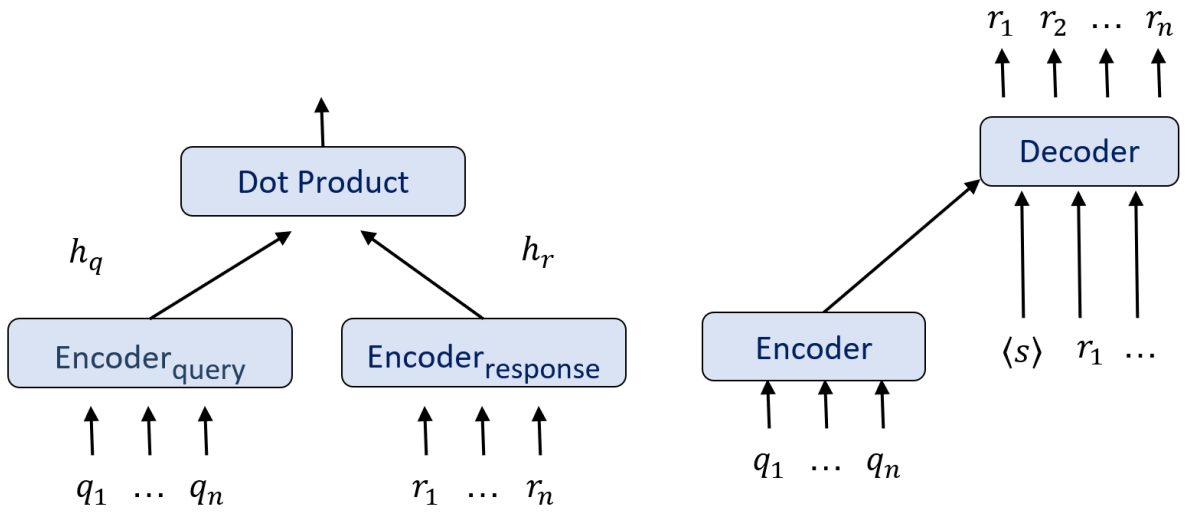

Figure 7.3 Retrieval- and Generation-based dialog system architecture (Jurafsky and Martin, 2023).

The End-to-End architecture unifies the modules for NLU, DM, and NLG and produces a system response either based on retrieval methods or generation methods (see Figure.7.3). The retrieval-based method selects a response from a dialogue corpus dependent on its appropriateness to the context. For this, a bi-encoder model is applied. One encoder is trained to generate a contextual embedding of the textual representation of the user's utterance. The other encoder is trained to produce an embedding of the candidate responses provided by large data corpora. The dot product between the two embedding vectors can then be used to calculate a similarity score. For determining a response, the candidate with the highest similarity is selected. The embeddings may be produced using different techniques, e.g., BERT or Word2Vec. For enhancing the quality of the selection, more

context, than solely the last user utterance, e.g., the dialogue history or information about the user's sentiment, may be included for producing the embeddings.

The generation method utilizes an encoder-decoder model for creating a system response. Thus, the problem of finding an appropriate system response given user input can be seen as a translation task. Here, the encoder network first creates a context vector that represents the user input and the dialogue history so far using BERT, for example. Afterward, the decoder network uses this vector to create an output considering the context vector and the response generated so far. Basic encoder-decoder models are prone to provide repetitive and dull responses. Therefore, some modifications are necessary to provide a more suitable system response. For example, by using reinforcement learning or adversarial networks for achieving a more natural conversation. ChatGPT makes use of a method called "Learning from Human Feedback" for providing more human-like responses based on human ratings of candidate responses (Ziegler et al., 2019).

4.3 Enhancing the Cognitive Abilities of Conversational Agents

Current architectures of dialog systems for conversational agents are highly performant for achieving specific tasks or conducting open-ended conversations using natural language. However, they are restricted as they are usually limited to solely finding appropriate responses to user input, while they do not reflect other fundamental aspects of human behavior during a conversation. For example, they do not adapt and personalize their responses to individual users or lack the ability to take proactive self-initiated actions. To achieve these kinds of behaviors, the cognitive capabilities of conversational agents need to be enhanced.

One important ability that needs to be included is to process multi-modal information. This means that conversational agents should be able to take in various types of information sources, not just speech, such as visual cues, physiological information, and context. By doing so, conversational agents can determine high-level user information, such as the user's affective state and emotion (Graesser et al., 2012), level of trust (Kraus et al., 2021), knowledge (Nothdurft et al., 2015), or satisfaction (Ultes et al., 2015). For example, combined audiovisual information can be used for estimating the emotional state of the user (Tzirakis et al., 2017). Here, audio and video data are first encoded in a multimodal representation amenable to computational processing. For this, LSTMS or convolutional neural networks (CNNs) can be used to extract a numerical representation of speech and visual features independently. Afterward, multimodal fusion techniques (Poria et al., 2017) are applied for joining features from both modalities to make predictions, e.g., by simply concatenating the feature vectors and feeding them into a neural network. The information about the user's emotional state can then be fed into a user model, which serves as a knowledge base for providing adaptive behavior. By modeling the user's behavior, characteristics, and goals, the conversational agent can adapt its

behavior to better meet the user's needs and expectations. For example, the user's emotional state can be used by a conversational agent to provide adequate emotional support by applying comforting strategies expressing empathetic and understanding behavior (Liu et al., 2021). Similarly, a conversational agent can provide multi-modal system output, e.g., utilizing gestures, facial expressions, and speech, for generating a more anthropomorphic user experience, which is often used in embodied conversational agents such as the GRETA agent (Pelachaud, 2017).

Furthermore, multi-modal information can be combined with advanced knowledge, reasoning, and planning abilities to achieve proactive behavior. For example, Kraus et al. (Kraus et al., 2020) developed a proactive conversational agent utilizing planning and ontological reasoning techniques for providing adequate support during the execution of DIY projects. Here, the user's task progress and their current activity with a modified electric screwdriver were tracked to initiate timely reflection dialogs. Evaluating their approach, the authors showed that the proactive agent was perceived as more trustworthy and led to higher user satisfaction with the project outcome than interacting with a reactive version of the agent. Further, proactive behavior can be used to adequately change topics in open-ended conversations with knowledge graph-based neural dialog systems. For example, Lei et al. (2022) used a reinforcement learning-based approach that includes task-relevant information, information about the user's previous satisfaction with the dialog, and the user's level of cooperativeness to determine the topic in the next dialog step. The goal was to achieve both fast task completion and user satisfaction.

Despite the high potential of augmenting conversational agents with multi-modal abilities, their application in commercial settings is quite limited. One reason is the sensitivity of user-specific data such as emotions, gender, age, or knowledge which raises several ethical and privacy questions. Further, high-level user information such as emotions and trusting behavior are quite subjective and differ greatly from individual to individual which limits the reliability of such recognition software for real-world application. Thus, adequate adaptation to recognized user states may fail which can result in poor performance and customer satisfaction as well as low acceptance of the applied systems. Therefore, multi-modal user information needs to be handled carefully and their reliable and safe application is still future work.

5 The social dimension

5.1 Social cues

In the field of HCI, Nass and his colleagues proposed the Computers Are Social Actors (CASA) paradigm and demonstrated that people mindlessly apply social scripts from human-human interaction when

they use computers (Nass et al., 1994; Reeves & Nass, 1996). However, technological advancements and artificial developments have led to a life in which we are surrounded by media technologies, such as conversational agents. To account for these advancements, Lombard and Xu (2021)proposed a structural extension of the CASA paradigm, i.e., the Media are Social Actors (MASA) paradigm. According to the MASA paradigm, social cues are triggers of users' social responses to media technologies. Social cues are physical or behavioral features displayed by a conversational agent which are salient to users (Fiore et al., 2013). An example of a social cue in conversational agents is an avatar which can be more or less human-like. Subsequently, social cues are (sub)consciously interpreted by users in the form of attributions of mental state or attitudes towards the conversational agent (Fiore et al., 2013; Wiltshire et al., 2014). For example, a conversational agent with a machine-like avatar might be interpreted as impersonal, which in turn might lead to the users' social response of attributing less trust in its capabilities.

Chaves and Gerosa (Chaves & Gerosa, 2020) distinguish several social cues chatbot designers can implement in chatbots: conversational intelligence refers to the chatbot's ability to manage interactions with users (for example by proactively sending relevant messages or asking follow-up questions to the user), social intelligence describes the chatbot's impact on the social behavior of the user (the chatbot could, for example, evoke appropriate or inappropriate behavior from users), and personification includes chatbot characteristics that make the agent appear more humanlike (such as the usage of a human-like avatar and communication style). An overview of Chaves and Gerosa's (2021) social cues is shown in Table 7.2.

Chaves and Gerosa's social cues can be implemented in various phases of the communication journey users go through when having a conversation with a conversational agent. Users can have expectations about the conversational agent before the actual conversation (phase 1), thereafter they will interact with the conversational agent (phase 2), and subsequently, they will formulate overall evaluations of the agent and possibly also of the organization that the conversational agent represents (phase 3) (van der Goot et al., 2020).

In the following paragraphs, we will reason the desirability and necessity to include the social cues in the different phases of users' communication journey with the chatbot, how the cues can be operationalized, and their relative impact on the users' prior expectations of the chatbots, their responses during the conversation, and perceptions and behavioral intentions after the conversation. Moreover, we will illustrate the applicability of the social cues in each phase by discussing two chatbots from different domains: a customer service chatbot, and a smoking cessation chatbot.

Table 7.2: Overview of the social cues that can be used in conversational agents per chatbot journey phase and examples of operationalizations (phase 1 and 2) and effects (phase 3)

		Chatbot user journey		
Social cues	**Description**	**Phase 1:** **Prior to the interaction**	**Phase 2:** **During the interaction**	**Phase 3:** **After the interaction**
Conversational intelligence	The chatbot's ability to manage interactions with users			
Proactivity	The chatbot takes initiative autonomously, resulting in a two-way conversation	+- include a personal greeting, but avoid intrusiveness	+ ask follow-up questions, provide relevant information, monitor users' goals and guide users towards their goals using motivational messages.	+ could contribute to the user's task and to having a smooth conversation
Conscientiousness	The chatbot demonstrates attentiveness to the conversation	- only applicable if the chatbot can use conversation history at the start of a new conversation	+ Keep the user aware of the chatbot's context, provide meaningful answers, use and balance confirmation messages	+- could contribute to the user's task and to having a smooth conversation

Communicability	The chatbot conveys its features to users	++ explain the purpose of the chatbot and its functionalities, and mention the option to redirect users to a human agent	+- remind users about the purpose of the interaction, and redirect the users to a human agent in case of failure or conflicts	+ could contribute to having a smooth conversation
Social intelligence	The chatbot's impact on the social behavior of the user			
Damage control	The chatbot's ability to deal with failures and failures and conflicts	+ mark the chatbot's (in)competence, mention to option to be redirected to a human agent	++ avoid failures, recognize them, communicate failures properly, provide the option to be redirected to a human agent	++ could contribute to the user's task and to having a smooth conversation
Thoroughness	The chatbot's ability to be precise and consistent in language use and communication style	++ match the chatbot's language use and communication style to the context in which it is implemented	+ be consistent in communication style and language use, balance the granularity of the information	++ could contribute to relationship building
Manners	The chatbot's ability to show	+ include a personal	++ adopt speech acts, like	++ could contribute to

	polite behavior and adhere to conversational habits	greeting, self-introduction, and adhere to turn-taking protocols	opening and closing sentences, acknowledgements, make interactions personal	having a smooth conversation and to relationship building
Moral agency	The chatbot acts based on social notions of right and wrong	+ avoid stereotypes in the chatbot's avatar, name, and reference to the user	+ use 'clean' training data (without harassment), be aware of biases	+ could contribute to relationship building
Emotional intelligence	The chatbot recognizes users' feelings and demonstrates respect and understanding	+- include a personal greeting, show information from prior conversations, if possible	+ chatbot shows empathy, reciprocity, and conscientiousness	+ could contribute to relationship building
Personalization	The chatbot's ability to adapt the interface, content, and behavior to the users' preferences, needs, and situational context	++ build the chatbot on the basis of cultural, behavioral, personal, conversational, and contextual data	+- use information from the conversation to tailor the responses, but avoid intrusiveness	+ could contribute to relationship building
Personification	The chatbot's characteristics that make the agent appear more humanlike			

Identity	The chatbot's appearance and cultural traits	+- include a disclosure about the artificial nature of the chatbot. Also the avatar and name type can mark or mask the chatbot's artificial identity	+ elaborate on the chatbot's persona with a matching language style	+ could contribute to relationship building
Personality	The chatbot's behavioral traits	+ develop and introduce a clear personality trait	+ use appropriate language, have a sense of humor	+ could contribute to relationship building

Note. Meaning of the characters: ++ highly desirable / necessary, + advisable, +- moderately desirable because of pros and cons, - discourage.

.

5.2 User expectations prior to the conversation

The first phase of the users' communication journey with a conversational agent consists of the agent's first messages. The social cues that are present in these first messages could influence users' expectations about the chatbot. Consider the customer service chatbot of Cana Brava Resort[3] presented in Figure 7.4 (left). Although the term 'virtual attendant' and the tottler-like avatar implicitly disclose the interlocutor is not a human being but a conversational agent, several social characteristics are adopted that could stimulate social behaviors that are habitual in human-human conversations (Chaves & Gerosa, 2020; van Hooijdonk et al., 2023). The chatbot contains a humanlike name, avatar, and informal communication style that all contribute to the personification of the agent. These cues could enhance users' perceptions of the humanlikeness of the chatbot (Liebrecht et al., 2020). Also, the chatbot of Cana Brava Resort uses social intelligence cues: a personal greeting ('Hi'), a self-introduction ('I'm Jorginho!'), and turn-taking ('How can I help you today?'). These cues display or adhere to manners and hence correspond to social norms from human-human communication that could stimulate users to act in the same social intelligent way (Chaves & Gerosa, 2020).

However, several opportunities to shape user expectations even better remain untapped by the Cana Brava Resort chatbot. First of all, it can be questioned what expectations users will have about the chatbot's competence. The tottler-like avatar and its rather informal language style could on the one hand increase users' perceptions of warmth, but on the other hand also lower expectations regarding the chatbot's competence (Khadpe et al., 2020) which in turn negatively affects users' trust in the chatbot capabilities. In contrast, communicating the intelligence of the agent, for example by stating the chatbot is 'an expert' for hotel bookings, users' trust in the chatbot's capabilities could be increased. The implementation of communicability cues in a chatbot introduction could also positively affect users' intentions to engage in a conversation with a chatbot. The chatbot could manage expectations about its purpose to avoid that users would address topics on which the chatbot has not been trained ('I can help you with hotel bookings'), it could give users guidance on how to interact with the chatbot to avoid miscommunication ('please choose a topic or ask your question in a maximum of five words'), and it could reassure users that human agents are present if needed ('You always have the possibility to continue the conversation with a human agent') (van Hooijdonk et al., 2023). In contrast, the opening messages of the smoking cessation chatbot Quitly[4] in Figure 7.4 (right), contain multiple communicability cues by explaining the purpose of and interaction with the chatbot. It also shows good manners because it proactively greets the user personally ('Hey there, Christine').

[3] https://canabravaresort.com.br/

[4] https://m.me/quitly.bot

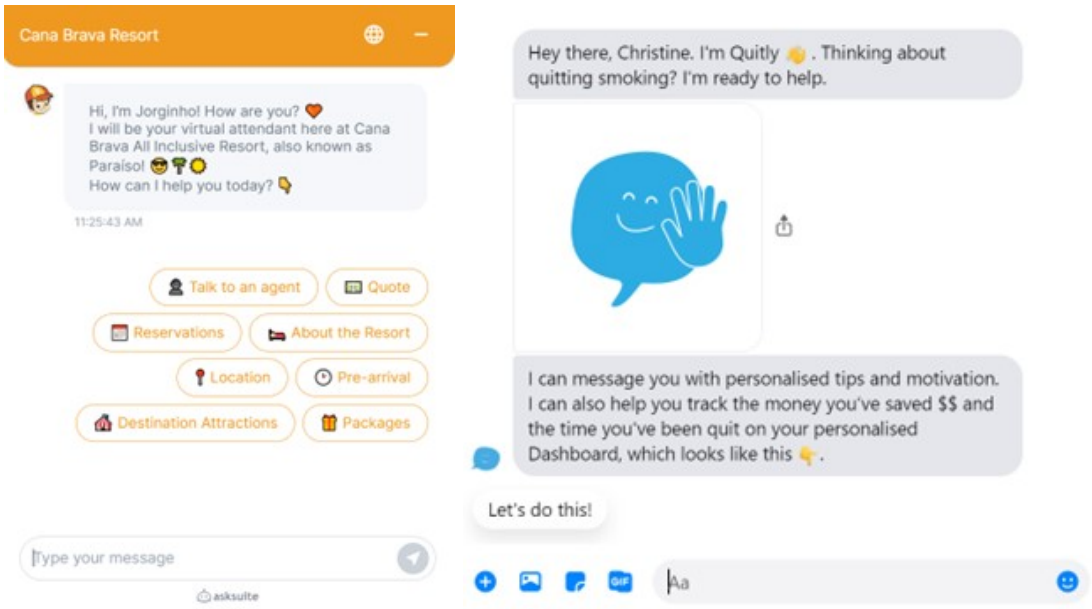

Figure 7.4: Introduction of 1) the customer service chatbot of Cana Brave Resort (left) and 2) Quitly the smoking cessation chatbot (right). Screenshots by the authors.

5.3 User experience during the interaction

During the interaction, the second phase of the user-chatbot communication journey, the conversational agent can display several social cues. It is important that these cues match the expectations users developed in the first phase of the journey. The customer service chatbot of Cana Brave Resort and the smoking cessation chatbot Quitly (see Figure 7.5) both adopt an informal and engaging communication style by using emoji, contractions, acknowledgments, and showing empathy. The customer service chatbot also communicates with a sense of humor ('I also love to eat well and have a few drinks! ;D'). Although this communication style can positively impact users' perceptions of the human likeness of the chatbot, Chaves and Gerosa (2020) argue that the chatbot's communication style - or even its personality - should be considered appropriate by its users. This depends on users' characteristics (e.g., age, literacy, and familiarity with chatbots), and context characteristics, such as the domain in which the chatbot is implemented (e.g., customer service vs. healthcare vs. politics), the organization that the chatbot represents (e.g., a formal banking company vs. an informal e-commerce company), and the chatbot's purpose (e.g., providing answers to FAQ's vs. providing psychotherapeutic support). A mismatch between users' expectations and perceptions about the chatbot's communication style could withhold users from continuing a conversation with the conversational agent.

Another challenge is the implementation of conversational intelligence cues proactivity and conscientiousness. Examples of the cues are personally greeting users by their names, using previously shared user information in conversations, or proactively suggesting new topics or follow-up questions.

Some of these cues can be observed in the smoking cessation chatbot, such as personal greetings and providing personalized tips (Figure 7.4 and 7.5). However, the implementation of these social cues also demands privacy protective operations, such as an explicit agreement from users that their personal information will be saved and reused at another point of time (which was the case when registering for Quitly). Thus, user and context characteristics should be taken into account in the implementation of conversational intelligence cues. These cues seem to be less relevant in the context of a task-based customer service chatbot which customers use to efficiently find a tailored response to their questions. However, these cues are more relevant in a socio-oriented context in which the chatbot is used as a coach or friend for a longer period of time.

Lastly, how the conversational agent deals with miscommunication affects users' experiences. Different social cues can be implemented to avoid and/or deal with miscommunication in a socially acceptable manner (Chaves & Gerosa, 2020). Miscommunication can be avoided by using conscientiousness cues, such as keeping the conversation on track by reminding users about the purpose of the interaction and informing them about the next steps. Moreover, confirmation messages can be used to check the chatbot's understanding of user messages. These conscientiousness cues are especially important in task-based customer service chatbots which customers use to reach a goal in an efficient and productive way (Chaves & Gerosa, 2020; Duijst, 2017). Miscommunication often occurs due to lack of chatbots' linguistic and world knowledge (Wallis & Norling, 2005). Social intelligence cues can be used to deal with miscommunication, such as recognizing and apologizing for it, which may reduce users' annoyance and frustration. Also, communicability cues can be used to recover miscommunication, such as providing options (e.g., 'Sorry, I did not understand you. Do you want to know more about 1) beverages or 2) breakfast?'), or redirecting the users to a customer service employee. The Cana Brava Resort chatbot Jorginho and smoking cessation chatbot Quitly adopt different approaches to deal with miscommunication which occurred at the start of the conversation. Chatbot Jorginho does not recognize the miscommunication explicitly. Instead, it gives the user directions on how to communicate with the system by means of the open text field and by showing the buttons that hints the user towards the topics that the chatbot can handle (see Figure 7.6 (left)). Chatbot Quitly, in contrast, explicitly states that it was unable to understand the user's utterance, and subsequently shows options to steer the conversation in the right direction (see Figure 7.6 (right)).

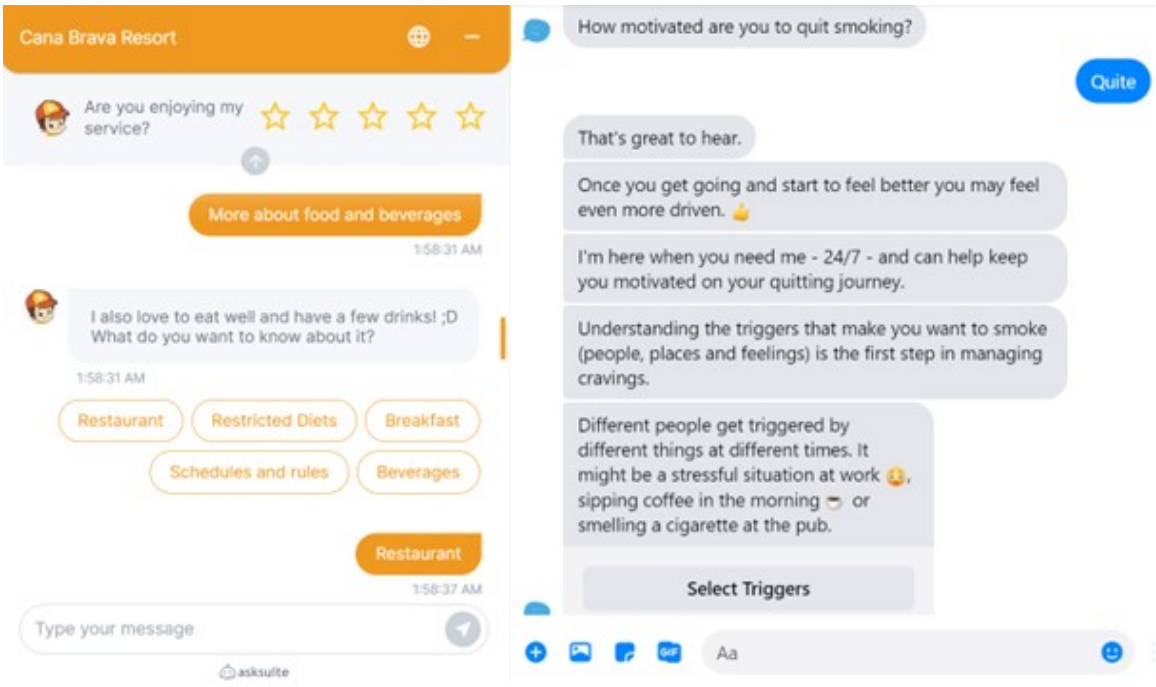

Figure 7.5: Conversation with 1) the customer service chatbot (left) and 2) the smoking cessation chatbot (right). Screenshots by the authors.

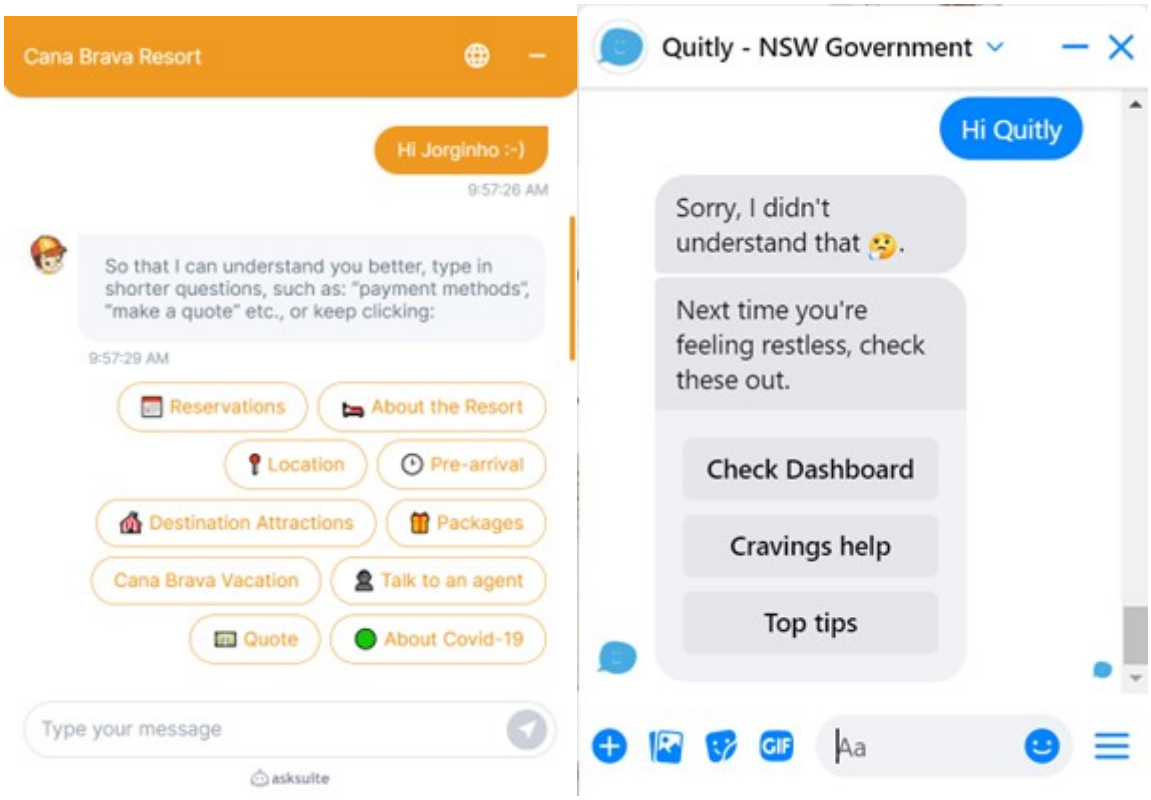

Figure 7.6: Miscommunication with 1) the customer service chatbot (left) and 2) the smoking cessation chatbot (right). Screenshots by the authors.

5.4 User evaluation after the interaction

In the third phase of the chatbot user journey, users form an overall evaluation based on their experiences with the conversational agent. Interestingly, this phase differs amongst the customer service chatbot and the smoking cessation chatbot. The customer service chatbot of Cana Brava Resort is a task-based chatbot that users visit to make a booking or to obtain more information about the resort. When the user approaches the end of the conversation, the chatbot asks 'Was I able to answer you?'. In case the user responds 'yes', a positive experience and evaluation of the chatbot could motivate them to start a new conversation with the agent next time they want to make a booking or obtain more information about the resort (see Figure 7.7). In contrast, the conversation of the smoking cessation chatbot is designed in such a way that there is no endpoint in the conversation (see Figure 7.8). It is therefore important that user evaluations maintain high to stimulate them to keep using or re-using the chatbot again.

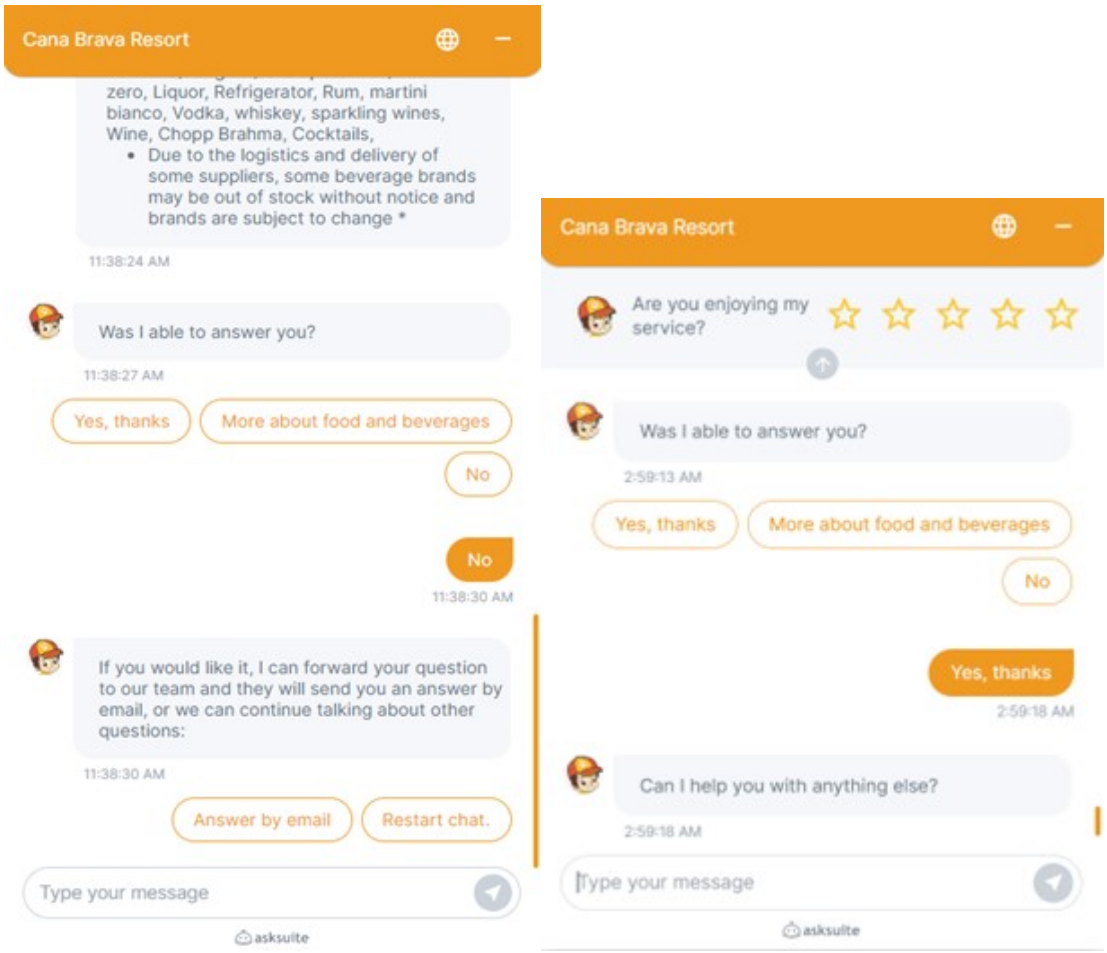

Figure 7.7: Final stage of the conversation with the customer service chatbot of the Cana Brava Resort. Screenshot by the authors.

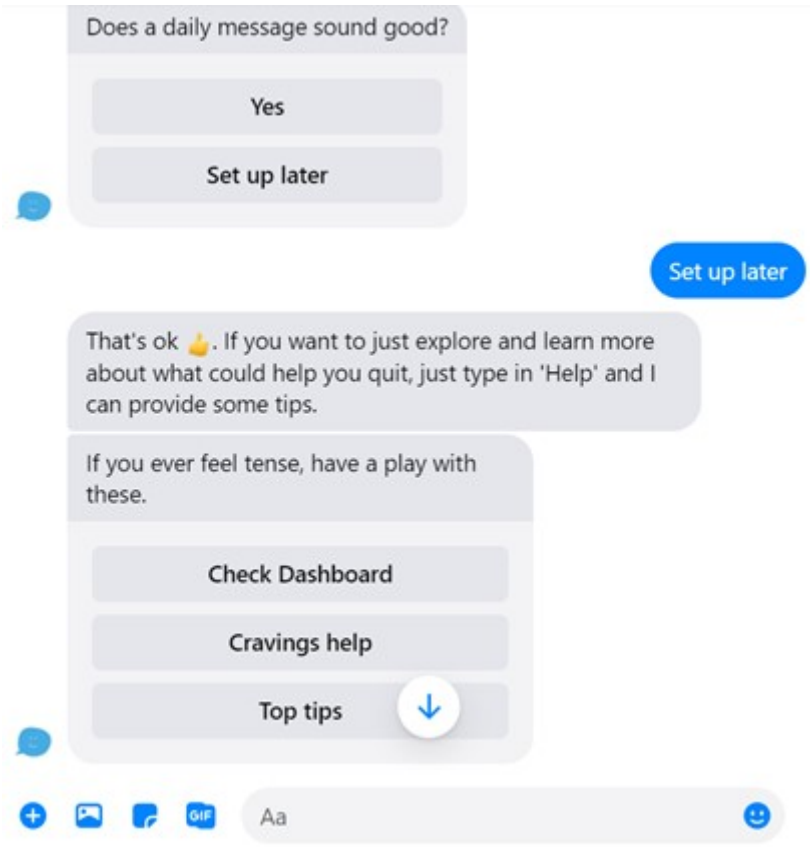

Figure 7.8: The conversation with the smoking cessation chatbot does not contain a clearly defined final stage of the conversation. Screenshot by the authors.

These experiences are related to the three goals users have when they engage in a conversation with a conversational agent. The first goal concerns the task users want to perform with the conversational agent, such as booking a hotel or learning more about smoking triggers (Luger & Sellen, 2016; Shechtman & Horowitz, 2003). Miscommunication may lead to an unsuccessful conversation resulting in users not achieving their tasks. This will lead to a negative evaluation of the interaction, the chatbot, and the organization. Therefore, miscommunication should be prevented in the first place, and recognized and solved in the second place. Miscommunication can be prevented if users' expectations about the chatbot's purpose and capabilities were managed before and during the interaction by implementing communicability cues. These cues could positively impact users' perceptions about the chatbot's capabilities (Jain et al., 2018) as well as trust and satisfaction. Moreover, research by Ashktorab et al. (2019) shows that users favored a chatbot that acknowledges miscommunication, and provides options of possible intents. This way the chatbot initiates the repair of the miscommunication, corresponding with the general preference for self-repair in interpersonal communication (Schegloff et al., 1977) and steers the conversation in a direction within its capabilities.

The second goal users have when engaging in a conversation with a chatbot is how they can have a smooth conversation with it (Luger & Sellen, 2016; Shechtman & Horowitz, 2003) which can be achieved by implementing conversational intelligence cues. For example, the chatbot takes initiative autonomously and starts a two-way conversation (proactivity cues), or the chatbot demonstrates attentiveness to the conversation (conscientiousness cues). These conversational intelligence cues are more difficult to implement as they have to be realized in the architecture of the conversational agent. Advances in AI technology might lead to the implementation of conversational intelligence cues enabling conversational agents to interact with users in a more intuitive way (Huang & Rust, 2018).

Finally, the third user goal refers to maintaining a certain relationship with the conversational agent (Luger & Sellen, 2016; Shechtman & Horowitz, 2003) which can be achieved by implementing social cues that simulate perceptions of humanlikeness, such as personification cues. This can be achieved by creating a chatbot with a clear personality trait, and by using identity cues, such as a humanlike name and avatar, and a personal and engaging communication style. Research shows that the usage of humanlike cues increases perceptions of humanlikeness which in turn leads to a positive evaluation of the organization (Araujo, 2018; Go & Sundar, 2019; Liebrecht et al., 2020). In order to hold these effects, Chaves and Gerosa (2020) state that the chatbot should maintain its personality and identity throughout the whole conversation (thoroughness cue), provided that the chatbot personification matches the user and the context (personalization cue). As these social cues are relatively easy to implement in conversational agents, a majority of current conversational agents contain (some of) these social cues, which has also been shown by the customer service chatbot and the smoking cessation chatbot in Figures 7.4 and 7.5. However, a conversational agent merely mimics humanlike communication. In order to maintain a relationship with a conversational agent it is important that it recognizes users' feelings and demonstrates respect and understanding. One way how these emotional intelligence cues can be implemented is by automatically analyzing the user's language, or by explicitly asking users how they feel. The latter approach can be observed in current conversational agents. Advances in AI technology might lead to the implementation of emotional intelligence cues enabling conversational agents to interact with users in a more empathetic way (Huang & Rust, 2018).

6 Conversational agent evaluation

The previous sections underline conversational agent design dimensions, including technical, social, and conversational capabilities. However, to foster the adoption and acceptance of these technologies, it is crucial to ensure that these dimensions are well-designed, meet user social and emotional expectations, and measure up to the desired functional performance and purpose.

Many commercially available conversational agents apply marketing strategies to evaluate the interaction between the software and their customers. These strategies include satisfaction surveys

or prompts for feedback. For example, in 7.6, the Cana Brava Resort chatbot, introduced in Section 7.5, prompts customers to evaluate the provided service using five-star ratings. These prompts are insightful to the companies as they assess the impact of the technology adoption on their target consumer's perceptions. However, to get an in-depth understanding of human-conversational agent interactions, it is crucial to have broader investigations to create guidelines that can be generalized to a variety of conversational agent's contexts of use.

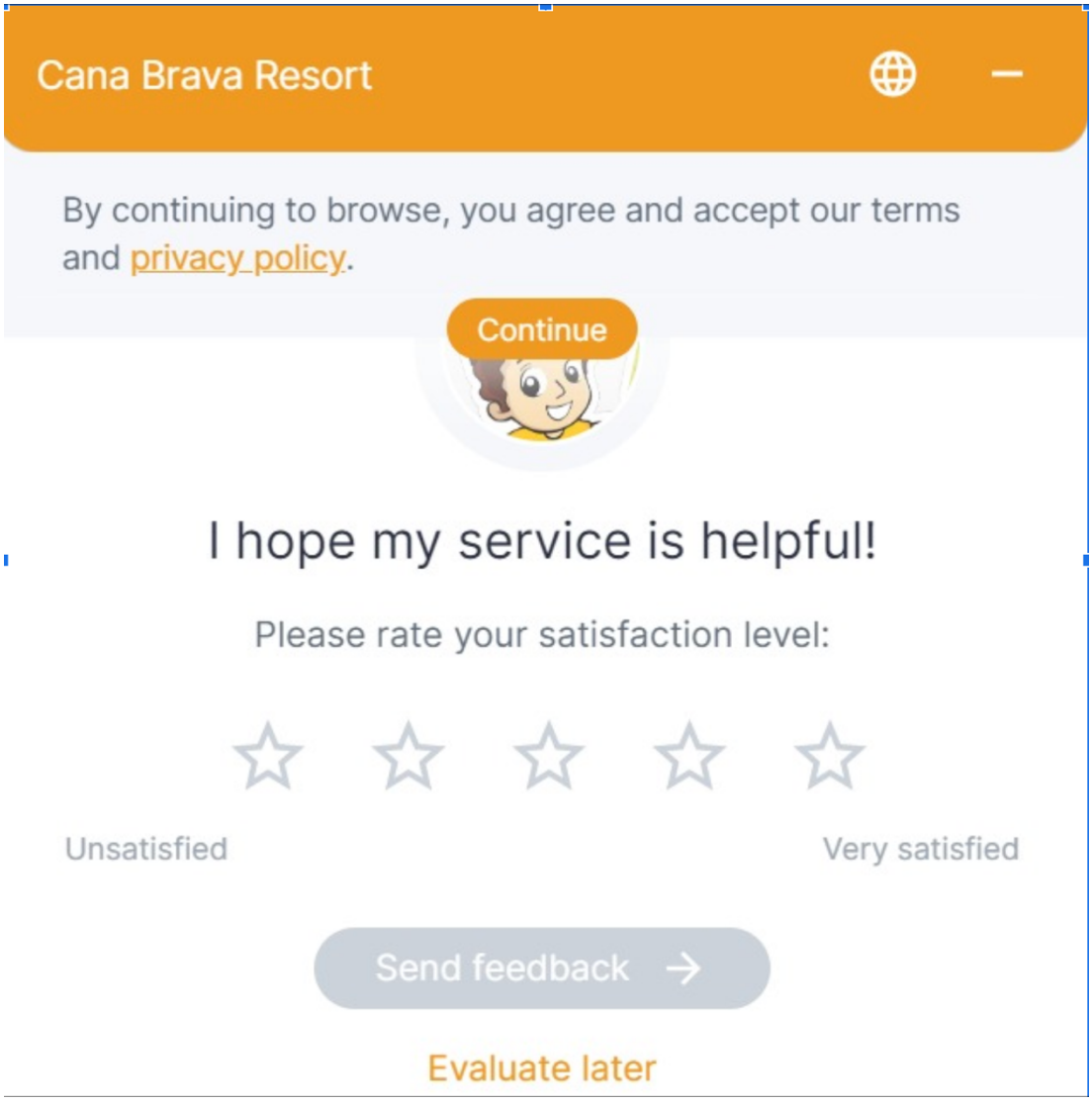

Figure 7.9: Cana Brava Resort chatbot asks for customer's feedback using five-star ratings. Screenshot by the authors.

The HCI field provides a set of well-established techniques to evaluate human-computer interactions, such as usability and User eXperience (UX) evaluation methods. However, these techniques were developed before the exponential spread of conversational interfaces; hence, they were designed and tested in the context of human-computer interactions over traditional interfaces.

Conversational agents, however, have their own needs in terms of evaluation. For example, conversational agent designers are more likely to prioritize linguistic style, manner, and perceptions of humanity than traditional interface designers. As anthropomorphic cues such as identity and use of natural language are more evident for these agents than for traditional interfaces, aspects such as satisfaction, trust, and engagement may have a different meaning and relevance from the user's perspective. Additionally, conversational agent design must care about interaction flow and context differently than traditional interfaces. Thus, evaluating a conversational interface may require developing new evaluation tools or adaptations to well-established techniques.

In the last decade, scholars in the HCI field have steered the gear toward assessing the extent to which currently available evaluation methods fit modern conversational interactions. Additionally, there have been efforts to validate new or adapted evaluation tools. This section presents pragmatic and hedonic aspects that determine user experience and acceptance of conversational agent technologies and examines methodologies commonly used in the conversational agent's context.

6.1 Usability and user experience

User experience (UX) is a comprehensive concept that refers to a person's perceptions and behaviors while using a software product (ISO 9241-11, 2018). Due to the concept's intrinsic complexity, a consolidated evaluation method to encapsulate all aspects of UX is yet to be developed. Instead, various evaluation methods have been used to address particular aspects of user experience, mainly distinguishing between pragmatic and hedonic goals (Bevan, 2009) associated with the user experience.

The evaluation of conversational agents follows the same practices. On the one hand, conversational agents are designed and evaluated to reach usability goals, such as effectiveness and efficiency. On the other hand, understanding the user's emotional experiences is crucial to achieving acceptance and adoption, which calls for assessing hedonic qualities such as engagement and pleasure (Haugeland et al., 2022).

Various well-established instruments have been used to evaluate conversational agents when focusing on usability metrics. For example, Guerino and Valentim (2020) mapped the literature to identify technologies used to assess the usability and UX of voice-based conversational systems. Unsurprisingly, the authors list technologies such as the System Usability Scale (SUS), the NASA Task Load Index (NASA-TLX), the Computer System Usability Questionnaire (CSUQ), and Nielsen's heuristics. Ren et al. (2019) found similar results in the context of chatbot evaluations. Table 7.3 shows some generic tools used to evaluate the usability and UX of conversational agents, returned by Guerino and Valentim (2020) and Ren et al. (2019).

Table 7.3: Evaluation tools used to evaluate the usability and UX of conversational agents. Adapted from (Guerino & Valentim, 2020; Ren et al., 2019)

Evaluation tool	Description
System Usability Scale (SUS)	Tool for measuring the Usability of products/services. It is a 10-item questionnaire that can be answered using a 5-point Likert scale, where 1 represents scale, where 1 represents "strongly disagree" and 5 represents "strongly agree". Even though it is not specific for conversational agents, SUS is a consolidated technology in Usability evaluation. Moreover, SUS is widely used in the literature to evaluate efficiency, effectiveness, and user satisfaction.
NASA Task Load Index (NASA-TLX)	NASA-TLX is a subjective assessment tool used to rate the perceived workload to evaluate a task or system. The aspects evaluated by this tool are mental demand, physical demand, time demand, performance, effort, and frustration.
Computer System Usability Questionnaire (CSUQ)	CSUQ is a 19-item questionnaire to be answered on a 7-point Likert scale. The questionnaire assesses utility, information quality, interface quality, and overall usability.
Attrakdiff	It is an instrument that evaluates the product's attractiveness in terms of usability and appearance. Attrakdiff contains 28 items of opposite pairs that can be answered on a 7-point scale.
Input Device Usability Questionnaire (IDU)	It is a 15-item questionnaire designed to investigate user interaction, distraction, ease of use, user comfort, frustration, enjoyment, error correction, and general usability. The questionnaire can be answered using a 5-point Likert scale.

However, both literature reviews (G. C. Guerino & Valentim, 2020; Ren et al., 2019) highlight that, in many cases, researchers create evaluation tools tailored to the particular study, which suggests the need for more specific evaluation methods. For instance, the most common tool reported by Guerino and Valentim was the creation of questionnaire-based assessment instruments for a specific study or agent. However, this ad-hoc creation, despite generally allowing a pertinent evaluation of Usability or UX, does not allow the replicability of the instrument in other studies or conversational agents.

Therefore, there have been initiatives to adapt existing evaluation tools to the context of conversational agents. Langevin et al. (2021) extended Nielsen's heuristics for usability to be used in the formative evaluation of conversational agents. The authors found that evaluators identified more usability issues when using the adapted heuristics when compared to the generic Nielsen's heuristics. According to the authors, conversational agent heuristics can be generalized to text-based, voice-based, and multi-modal conversational agents.

In the context of chatbots, Borsci et al. (2022) developed the Bot Usability Scale (BUS-15), which consists of a 15-item questionnaire focused on the user's satisfaction with the interaction. The questionnaire includes items about the chatbot's reachability, functionalities, quality of conversation and information, privacy and security, and response time. However, the scale has not been extensively tested and validated, requiring a collective effort from the research community to assess the reliability of the instrument.

When focusing on voice-based interfaces, Guerino et al. (2021) developed an instrument to evaluate Usability and UX, called U2XECS. The instrument is questionnaire-based, and from its application, it is possible to evaluate the following aspects: User Satisfaction, Efficiency, Effectiveness, Generic UX, Affect/Emotion, Enjoyment/Fun, Aesthetics/Appeal, Engagement/Flow, and Motivation. U2XECS was evaluated in controlled experiments; however, further studies are needed to generalize its results and enhance its benefits for the conversational agents' community.

However, there is more to the user experience than usability and satisfaction. Bevan (Bevan, 2009) argues that user experience in user-centered design also includes understanding what users do and why when interacting with the systems and maximizing the achievement of hedonic goals. Guerino and Valentim (2020) show that there is no clear distinction, on the part of evaluators, about what is included when evaluating UX; in this sense, it was seen that the same tool used to evaluate UX in one study was used to evaluate usability in another. Thus, more studies are needed so that, in addition to better understanding the definitions and aspects of UX and usability in conversational agents, specific tools are proposed for this context and replicable to other studies.

6.2 User perceptions and acceptance

The success of software or technology strongly depends upon the user's perceptions of the system,

such as expectation, engagement, and trust (Venkatesh et al., 2011). Combined with the usability and satisfaction dimensions of UX, positive user perceptions increase the likelihood of acceptance and adoption of a system. In the conversational agents' field, user perceptions have a unique characteristic when compared to other systems: the anthropomorphisation triggered by the use of natural language and the social implications associated with it (see Section 7.5).

Standardized questionnaires fail to capture the nuances of subjective user perceptions. Therefore, user perceptions are often measured using qualitative methods such as interviews and focus groups based on the interaction with a proposed conversational agent or the user's previous experiences with the technology (Bevan, 2009).

One technique that supports the qualitative evaluation of conversational agents is the Wizard of Oz (WoZ) (Dahlbäck et al., 1993) (see more in Section 7.3). This technique helps evaluators to consider the unique qualities of human-conversational agent communication in the initial design stages. Besides, this method enables control for implementation limitations that could harm usability and performance thus negatively affecting the user experience. As a drawback, some human limitations are difficult to overcome in WoZ studies. For example, response time can be difficult to simulate, as humans will unlikely be able to answer user inputs as fast as a conversational agent. Nevertheless, the benefits of the WoZ technique surpass its limitations, and the popularity of the technique has even fostered the development of environments to support its use for conversational agent evaluation (Simpson et al., 2022).

When the conversational agent evaluation is based on a current interaction, a common method is to invite the subjects to Think Aloud. Think Aloud sessions aim to identify user perceptions based on what they say out aloud about what they are doing and why. According to Barbosa et al. (Barbosa et al., 2022), when the user verbalizes their thoughts during the interaction, evaluators can capture genuine reactions and interpret problematic parts of the interaction.

Unlike user perceptions, which are a broad set of behaviors and emotions triggered by the interaction with conversational agents, acceptance is a more consolidated construct. Therefore, the literature provides standardized evaluation tools to measure the acceptance of a software system. The two most frequently used theoretical frameworks to evaluate the adoption of conversational agents are the Technology Acceptance Model (TAM) and the Unified Theory of Acceptance and Use of Technology (UTAUT2). However, both questionnaires are highly oriented to pragmatic qualities, such as usefulness and effort expectancy.

Moreover, by applying UTAUT2 to evaluate the acceptance of a healthcare-specific conversational agent, Laumer et al. (2019) found that using a specific technology (conversational agent) can influence the use of a particular instance of this technology (e.g., a conversational agent for disease diagnosis). The authors foster further research that proposes measurement items of acceptance specific to

application contexts. Furthermore, the authors extend the conclusion made by Venkatesh et al. (2011) by mentioning that it is necessary to investigate the influence of environmental factors on the acceptance of a conversational agent and to propose new models to measure this influence and its relationship with other factors empirically.

In this sense, Ling et al. (Ling et al., 2021) surveyed the literature to identify the factors that influence users' adoption of conversational agents. As a result, the authors propose a model where acceptance is driven by the usage benefits, which translates to the value gained from interacting with the agent. These benefits are influenced by both agent and user characteristics, including, for example, the agent's appearance and anthropomorphic cues and the user's demographic characteristics and intrinsic motivation. However, the proposed collective model is theoretical and there is still a lack of a practical framework to guide the use of the model for conversational agent evaluation.

7 Conversational Agents for social good

This section focuses on the impact of introducing conversational agents in our society, particularly focusing on the role these agents may perform in well-being and inclusion. Given the interest in conversational agents for social good, be they chatbots (Følstad et al., 2018), AI systems (Floridi et al., 2020), and/or robots (H. R. Lee et al., 2019), we first dive into what "social good" means.

We then discuss relevant application domains in which conversational agents serve as a tool for social good.

We relate social good to people's capacity to lead a "good life", such as the Aristotelian sense of eudaimonic well-being (Aristotle, 2019). Additionally, social good considers opportunities to foster people's capabilities, such as opportunities to achieve one's health (Robeyns, 2009) or promoting people's values in everyday technologies (Friedman, 1996; Wambsganss et al., 2021). Thus, one broad definition of technologies for "social good" is to promote the well-being of people and nature while mitigating harm towards them (Floridi et al., 2020); harms can be due to data manipulation, e.g., incorrectly labeled training data for conversational agents, and violation of users' privacy, e.g., personally identifiable information from conversational data being shared with third parties without consent (Floridi et al., 2020). For these potential harms, what is needed is a clear explanation about how conversational agents share and handle personal data, as well as how they have been trained, e.g., on what dataset and how (e.g., Mitchell et al. (2019)), which can assist in expectation management between users, conversational agents, and other stakeholders. Another harm is systematically and continuously marginalizing specific target user groups by not involving them in the design and development process, e.g., designing conversational agents *for* people with disabilities rather than designing *with* them (e.g., Spiel et al. (2020)). Directly involving people who are intended users (especially from marginalized or potentially marginalized groups) in the design of these

conversational agents can mitigate potential harm, e.g., through value-sensitive design (see Section 7.3), capability-sensitive design (Jacobs, 2020), and participatory design (Frauenberger et al., 2017).

We turn to how conversational agents can serve the social good in specific domains, such as promoting well-being in the healthcare sector. Conversational agents are being utilized for physical and mental healthcare, such as assisting doctors or patient support (Preum et al., 2021). There are healthcare conversational agents that are for specific target populations, such as people with neurodevelopmental disorders (Catania et al., 2023) or adolescents (Fitzpatrick et al., 2017; Gabrielli et al., 2020), and again, it is important to bear in mind involving these intended user groups when developing such agents (e.g., Lopatovska et al. (2022)).

While well-being technologies come in many types, such as in VR or mobile apps (Calvo & Peters, 2014), what is particular about conversational agents is that they can refer to themselves in the first person, talking to human interactants as another "I", so instead of clicking or scrolling through apps, people are able to converse with technology (M. Lee & Contreras, 2023). Relatedly, conversational agents can make it easier for some to discuss sensitive topics like PTSD (compared to disclosing this to other people) as these agents are seen as "mere machines'' (Lucas et al., 2014, 2017). People may believe that machines or chatbots are less judgemental than other people (Brandtzaeg & Følstad, 2017, 2018). Besides targeted health interventions like sleep management (Rick et al., 2019) or smoking cessation (see Section 7.5 and (Perski et al., 2019)), there are many examples of conversational agents that support people's mental health (Ahmed et al., 2023), such as Wysa that is empathy based (Inkster et al., 2018), Woebot for depression (Fitzpatrick et al., 2017), and Vincent (M. Lee, Ackermans, et al., 2019) that asked for people's advice after reporting on its mishaps (Figure 7.10) which helped people be self-compassionate by being compassionate to it. Hence, conversational agents can be perceived to be helpful or in need of help in order to elicit specific psychological responses such as the tendency for self-disclosure or care-oriented behavior.

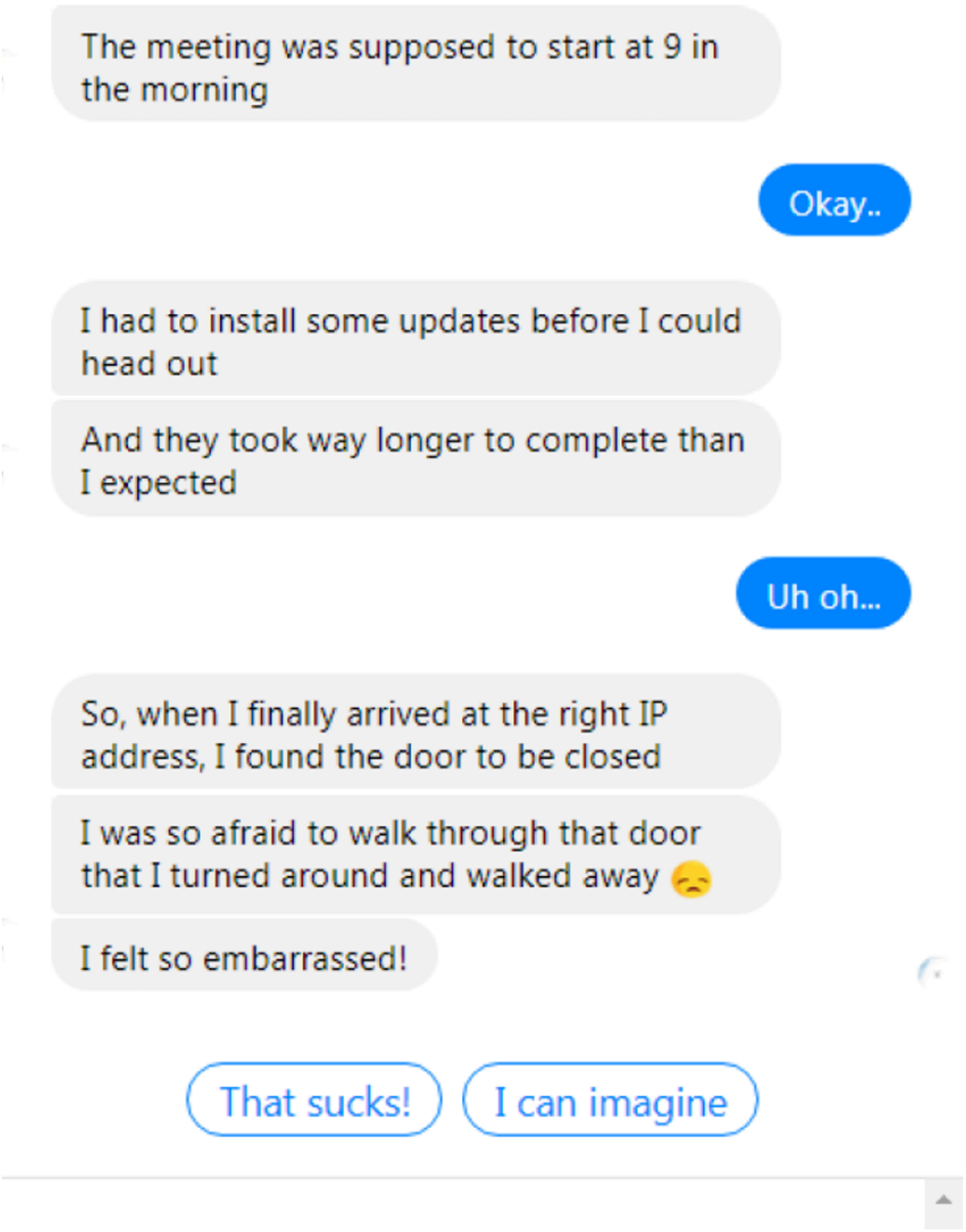

Figure 7.10. Vincent, a chatbot, sharing that it was late for a meeting to ask for help from people. Source: Lee et al. (2019).

One underlooked aspect in healthcare is any potential conflict between stakeholders; there might be differences in opinions between experts, healthcare providers, and patients, such as a positive perspective by the experts who believe in predictive health (De Maeyer & Markopoulos, 2021) vs. patients who may not want predictive health forecasting that conversational agents can share with them, especially in emerging forms like digital twins at home (De Maeyer & Lee, 2022). In what ways conversational agents should navigate across potential conflicts between stakeholders, as well as

being entangled in healthcare conflicts, will become increasingly more significant to address.

Conversational agents for well-being focus on individuals' health, but there are other social good areas that conversational agents can be active in, be it for education, public health crisis, personal finance knowledge, or for inclusive citizenship, e.g., government agencies using conversational agents to increase citizens' awareness of policies. Conversational agents used for educational purposes come in many varieties (Chen et al., 2023; Hwang & Chang, 2021; Khosrawi-Rad et al., 2022), such as agents that help foster creativity and curiosity (Abdelghani et al., 2022) or agents that assist in online learning via text and voice input in a MOOC (Winkler et al., 2020), e.g., by fostering inclusive learning through an active FAQ chatbot (Han & Lee, 2022). Learning can occur in workplaces, e.g., for reflection and journaling (Kocielnik et al., 2018). Segmenting needs and goals per intended group is important here; college students vs. children would have different learning challenges, for instance. Design features of conversational agents should consider various differences. For example, a simple, clickable button to start the conversational agent might be more appropriate for children rather than expecting them to use "wake words" (like "Hey Google") to start a voice-based agent (Catania et al., 2020) and children might learn better when the agent is portrayed as a peer than a tutor, i.e., due to higher engagement and attention to a peer agent (Zaga et al., 2015). Conversational agents can thus play diverse roles, such as being a mentor or assistants, as well as being a peer, student, tutor, or a teacher (Wollny et al., 2021). As with potential conflicts among stakeholders in healthcare, various perspectives have to be considered, e.g., differences in what children vs. parents may educationally value and how then conversational agents should be used at home (Garg et al., 2022; Garg & Sengupta, 2020), scaling up to potential differences in how schools, universities, or government agencies may want to institute conversational agents for education compared to students and parents. Conversational agents themselves can be the cause of conflict, such as a recent worry with students' usage of ChatGPT over the kinds of influence conversational agents can have on learners and if and how that should be curtailed (Rudolph et al., 2023).

Beyond education, there are other cases in which allowing people to help conversational agents would be helpful, e.g., humans teaching conversational agents classification tasks which can foster trust between humans and agents on crowd work platforms, which increases the agents' ability to perform tasks when they function as "teachable machines" that can grow with people they learn with (Chhibber et al., 2022). Community- or crowd-driven training also seems promising for a chatbot to provide on-demand therapy based on crowd workers training it (T. Abbas et al., 2020). Also on gaming platforms, people can teach a chatbot how to act as a community member, e.g., how to address people in the same community, which can strengthen behavioral norms that an online community wants to promote via teaching a chatbot (Seering et al., 2020). In general, this forms empowerment through a conversational agent perspective; people help their communities or themselves by

empowering an agent, e.g., learning through teaching (Tanaka & Matsuzoe, 2012) or caring for oneself by caring for an agent (M. Lee, Ackermans, et al., 2019).

Conversational agents are deployed for many other domains we briefly go over. Concerning a public health crisis, during Covid-19, many chatbots were deployed, e.g., to provide timely information on the spread of the virus though there are concerns about the potential spread of inaccurate data or misinformation (Miner et al., 2020). There are efforts to better assist citizens through conversational agents in digital governance (N. Abbas et al., 2023), like for reaching those who may need social service support (Simonsen et al., 2020). In the finance sector, conversational agents can support people's financial decision-making (Wube et al., 2022), e.g., for small businesses to get loans (Candello, Grave, et al., 2022), or for people to get an understanding of the stock market (Sharma et al., 2021) or for exploring alternative currencies, like a chatbot for the cryptocurrency marketplace (M. Lee, Frank, et al., 2021). In sum, promising directions for social good in healthcare, education, crisis management, digital governance, and finance (among others) are in place for conversational agents.

7.1 Challenges and opportunities for social good

The main challenge is that the notion of "social good" deserves to be critically questioned, given that many corporate, social good ventures do not always come with harm mitigation (Kwet, 2019; Zembylas, 2023). Broadly, another point is that what may be "socially good" is not universal (Madianou, 2021). This relates to how, for instance, accounting for race and socio-economic differences in the design and deployment of conversational agents (e.g. Garg & Sengupta (2019)) requires a nuanced view on *what* is socially good for *whom* from *whose perspective*.

Hence, what further needs to be discussed are user representation and conversational agent representation in terms of *intersectionality* (cross-cutting many attributes such as race, gender, disability, class, and more), such as what traits (like name and avatar) of a conversational agent are designed as well as what groups are representative users of it. Intersectionality is not yet a focus for conversational agents (both in how they are designed and how they are received by intended users), but we see a growing need to consider intersectionality, especially in mitigating bias in the design and use of conversational agents (Ciston, 2019; M. Lee, Noortman, et al., 2021).

Sexism is compounded (in social and socio-technical environments) when considering that many smart speakers are positioned as female-sounding assistants that habituates users into ordering them around (Søndergaard & Hansen, 2018; Strengers et al., 2020; Strengers & Kennedy, 2021), but this also holds for ableism, e.g., when robot assistants are designed as able-bodied and digital colonialism, e.g., when non-native English speakers are less understood by smart speakers since specific "native" accents are prioritized (Wu et al., 2022). As another example, chatbots have difficulties discussing racism in general (Schlesinger et al., 2018) and a negative case was Microsoft's Tay, a chatbot on

Twitter, tweeted racist and homophobic remarks based on how it was trained (Wolf et al., 2017). Unintended harm is still harm, which means that the long-term research and development phase is ideal to lower the chances of harm caused by conversational agents, as well as refining what "good" can be offered by them according to the perspectives of people who are intended to be end users.

The emergent users from so-called "developing countries" will increasingly make up a large portion of conversational agents and other technologies, but Western technologists and developers' assumptions about their needs and desires often can be misguided or misunderstood (Arora, 2019). In India, for instance, users may not have a strong preference for human vs. AI-generated responses in smart speakers, but their combination (human and AI) can be valued (Ahire, 2022). Additionally, the context of use may differ; public-facing agents that are installed in urban areas can be used by many people (compared to only at home), such as a voice-based agent that was shared by many in an Indian slum (Pearson et al., 2019). Additionally, tackling societal issues like gender inequality, like a chatbot for addressing gender norms for Indian adolescence (Agarwal et al., 2021), may need to be specific to how gendered norms have cultural dimensions.

Before conversational agents can be adapted to diverse languages, cultures, and contexts, there are other limitations to keep in mind: Local contextualization (as per above) requires an intersectional mindset, but also conversational agents are not yet adept at handling bilingual or multilingual dialogue (Cihan et al., 2022) and more work needs to be done to account for disabilities or different patterns of expressing oneself, e.g., voice agents for those who stammer (Bleakley et al., 2022) and those who are deaf or hard of hearing (Glasser et al., 2020). While there are ongoing challenges in making conversational agents that are accessible to a wide array of people, we also see many opportunities that conversational agents offer, e.g., conversational agents that assist students with autism, dyslexia, and other conditions to learn more independently (Lister et al., 2020). Hence, what promotes inclusion for some may be an exclusion for others, such as voice- vs. text-based interaction for people who stammer vs. those who do not; what helps then is to explore diverse modalities and diverse user groups for context-based design and deployment.

Lastly, we see the need to mature how privacy practices can be handled by conversational agents; current issues include the lack of (or outdated) privacy policies in many voice-based applications (Edu et al., 2022; S. Liao et al., 2020). As many applications get added on top of platforms, such as Spotify connecting with Alexa, cross-platform privacy policies also need to be considered. While people live and collaborate with conversational agents at home, work, and beyond, research should also focus on conversational repair, e.g., agents repeating directions vs. providing alternative options (Ashktorab et al., 2019), or how people will handle mistakes from conversational agents, e.g., an inaccurate scheduling assistant (Kocielnik et al., 2019). Conflicts may arise when people underestimate the agent and overestimate their own knowledge (and vice versa); this can be difficult to mediate situationally,

such as unclear on when an agent should take over compared to humans(Schaffer et al., 2019). Here, norms we have on other humans, such as blaming people for mistakes, may not always carry over to conversational agents (M. Lee, Ruijten, et al., 2021), making it difficult to assess how and when to hold human and non-human agents accountable (Lima et al., 2022; Nyholm, 2018).

8 The future

The path for conversation with intelligent software systems is already established. From now on, we expect that conversing with software systems will increasingly take over traditional interfaces. Therefore, it is crucial to reflect on future challenges, such as privacy and the ethical implications of long-term interactions with conversational agents.

ChatGPT has recently brought to light several discussions that demonstrate the fragility of our society to the full integration of conversational agents in our daily activities. ChatGPT capabilities resulted in extensive ongoing discussions on whether generative AI should be listed as authors in academic publications, the limits of plagiarism, and the impact on education (Bowers, 2023; Rudolph et al., 2023; Ueda & Yamada, 2023). In the upcoming years, the growing popularity of home and Internet of Things devices will continue to bring conversational interfaces to people's homes as well as work and leisure environments. These agents are expected to be omnipresent, and any device that knows the user is fully synchronized and able to respond to the user's requests (Gentsch, 2019). Combining large language models and omnipresent devices may enable the future development of conversational agents as complex as Samantha, from the movie *Her* (see Section 7.1). How this will affect our daily lives is a question still to be answered. In any case, like in the ChatGPT case, new privacy and ethical issues will emerge and must be resolved as the technology evolves.

References

Abbas, N., Følstad, A., & Bjørkli, C. A. (2023). Chatbots as Part of Digital Government Service Provision–A User Perspective. *Chatbot Research and Design: 6th International Workshop, CONVERSATIONS 2022, Amsterdam, The Netherlands, November 22–23, 2022, Revised Selected Papers*, 66–82.

Abbas, T., Khan, V.-J., Gadiraju, U., & Markopoulos, P. (2020). Trainbot: A conversational interface to train crowd workers for delivering on-demand therapy. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *8*, 3–12.

Abdelghani, R., Oudeyer, P.-Y., Law, E., de Vulpillières, C., & Sauzéon, H. (2022). Conversational agents for fostering curiosity-driven learning in children. *International Journal of Human-Computer Studies*, *167*, 102887.

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine*

Learning with Applications, *2*, 100006. https://doi.org/10.1016/j.mlwa.2020.100006

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., & others. (2020). Towards a human-like open-domain chatbot. *ArXiv Preprint ArXiv:2001.09977*.

Agarwal, D., Agastya, A., Chaudhury, M., Dube, T., Jha, B., Khare, P., & Raghu, N. (2021). *Measuring effectiveness of chatbot to improve attitudes towards gender issues in underserved adolescent children in India*. Tech. Rep.). Cambridge, MA: Harvard Kennedy School.

Ahire, S. (2022). Designing a Smart Speaker for Emergent Users: Human Plus AI Response. *Proceedings of the 13th Indian Conference on Human Computer Interaction, 2022*, 67–72.

Ahmed, A., Hassan, A., Aziz, S., Abd-Alrazaq, A. A., Ali, N., Alzubaidi, M., Al-Thani, D., Elhusein, B., Siddig, M. A., Ahmed, M., & others. (2023). Chatbot features for anxiety and depression: A scoping review. *Health Informatics Journal*, *29*(1), 14604582221146720.

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

Aristotle, T., translated by Irwin. (2019). *Nicomachean ethics*. Hackett Publishing.

Arora, P. (2019). *The next billion users: Digital life beyond the West*. Harvard University Press.

Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI'19)*, 254:1-254:12.

Augustsson, M. (2019). *Talking to Everything: Conversational Interfaces and the Internet of Things in an office environment*.

Barbosa, M., Nakamura, W. T., Valle, P., Guerino, G. C., Finger, A. F., Lunardi, G. M., & Silva, W. (2022). UX of Chatbots: An Exploratory Study on Acceptance of User Experience Evaluation Methods. *ICEIS (2)*, 355–363.

Barth, F., Candello, H., Cavalin, P., & Pinhanez, C. (2020). Intentions, meanings, and whys: Designing content for voice-based conversational museum guides. *Proceedings of the 2nd Conference on Conversational User Interfaces*, 1–8.

Bergman, A. S., Abercrombie, G., Spruit, S., Hovy, D., Dinan, E., Boureau, Y.-L., & Rieser, V. (2022). Guiding the Release of Safer E2E Conversational AI through Value Sensitive Design. *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 39–52. https://aclanthology.org/2022.sigdial-1.4

Bevan, N. (2009). What is the difference between the purpose of usability and user experience evaluation methods. *Proceedings of the Workshop UXEM*, *9*(1), 1–4.

Bleakley, A., Rough, D., Roper, A., Lindsay, S., Porcheron, M., Lee, M., Nicholson, S. A., Cowan, B. R.,

& Clark, L. (2022). Exploring smart speaker user experience for people who stammer. *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, 1–10.

Blythe, M. (2014). Research through design fiction: Narrative in real and imaginary abstracts. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 703–712.

Bordes, A., Boureau, Y.-L., & Weston, J. (2017). Learning end-to-end goal-oriented dialog. *International Conference on Learning Representations*. https://openreview.net/forum?id=S1Bb3D5gg

Borsci, S., Malizia, A., Schmettow, M., van der Velde, F., Tariverdiyeva, G., Balaji, D., & Chamberlain, A. (2022). The Chatbot Usability Scale: The Design and Pilot of a Usability Scale for Interaction with AI-Based Conversational Agents. *Personal and Ubiquitous Computing*, *26*(1), 95–119. https://doi.org/10.1007/s00779-021-01582-9

Bowers, A. J. (2023). *Unpacking the Caveats of ChatGPT in Education: Addressing Bias, Representation, Authorship, and Plagiarism*.

Brandtzaeg, P. B., & Følstad, A. (2017). Why people use chatbots. *Internet Science: 4th International Conference, INSCI 2017, Thessaloniki, Greece, November 22-24, 2017, Proceedings 4*, 377–392.

Brandtzaeg, P. B., & Følstad, A. (2018). Chatbots: Changing user needs and motivations. *Interactions*, *25*(5), 38–43.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., & Gasic, M. (2018). MultiWOZ-A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 5016–5026.

Calvo, R. A., & Peters, D. (2014). *Positive computing: Technology for wellbeing and human potential*. MIT press.

Cambre, J., & Kulkarni, C. (2020). Methods and Tools for Prototyping Voice Interfaces. *Proceedings of the 2nd Conference on Conversational User Interfaces*, 1–4. https://doi.org/10.1145/3405755.3406148

Candello, H., Grave, M., Brazil, E., Ito, M., Alves de Brito Filho, A., & De Paula, R. (2022). How can AI leverage alternative criteria and suggest a better way to measure credit worthiness and economic growth? *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–4.

Candello, H., Pinhanez, C., Muller, M., & Wessel, M. (2022). Unveiling Practices of Customer Service Content Curators of Conversational Agents. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 348:1-348:33. https://doi.org/10.1145/3555768

Carrol, J. M. (1999). Five reasons for scenario-based design. *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. Hicss-32. Abstracts and Cd-Rom of Full Papers*, 11-pp.

Cassell, J. (2000). Embodied conversational interface agents. *Communications of the ACM*, *43*(4), 70–78.

Catania, F., Spitale, M., Cosentino, G., & Garzotto, F. (2020). What is the Best Action for Children to" Wake Up" and" Put to Sleep" a Conversational Agent? A Multi-Criteria Decision Analysis Approach. *Proceedings of the 2nd Conference on Conversational User Interfaces*, 1–10.

Catania, F., Spitale, M., & Garzotto, F. (2023). Conversational agents in therapeutic interventions for neurodevelopmental disorders: A survey. *ACM Computing Surveys*, *55*(10), 1–34.

Chaves, A. P., Egbert, J., Hocking, T., Doerry, E., & Gerosa, M. A. (2022). Chatbots language design: The influence of language variation on user experience with tourist assistant chatbots. *ACM Transactions on Computer-Human Interaction*, *29*(2), 13:1-13:38. https://doi.org/10.1145/3487193

Chaves, A. P., & Gerosa, M. A. (2018). Single or Multiple Conversational Agents? An Interactional Coherence Comparison. *ACM SIGCHI Conference on Human Factors in Computing Systems*, 191:1-191:13.

Chaves, A. P., & Gerosa, M. A. (2020). How Should My Chatbot Interact? A Survey on Social Characteristics in Human–Chatbot Interaction Design. *International Journal of Human–Computer Interaction*, *37*(8), 1–30. https://doi.org/10.1080/10447318.2020.1841438

Chen, Y., Jensen, S., Albert, L. J., Gupta, S., & Lee, T. (2023). Artificial intelligence (AI) student assistants in the classroom: Designing chatbots to support student success. *Information Systems Frontiers*, *25*(1), 161–182.

Chhibber, N., Goh, J., & Law, E. (2022). Teachable Conversational Agents for Crowdwork: Effects on Performance and Trust. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–21.

Cihan, H., Wu, Y., Peña, P., Edwards, J., & Cowan, B. (2022). Bilingual by default: Voice Assistants and the role of code-switching in creating a bilingual user experience. *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–4.

Ciston, S. (2019). Intersectional AI is essential: Polyvocal, multimodal, experimental methods to save artificial intelligence. *Journal of Science and Technology of the Arts*, *11*(2), 3–8.

Core, M. G., & Allen, J. (1997). Coding dialogs with the DAMSL annotation scheme. *AAAI Fall Symposium on Communicative Action in Humans and Machines*, *56*, 28–35.

Cui, L., Huang, S., Wei, F., Tan, C., Duan, C., & Zhou, M. (2017). SuperAgent: A Customer Service Chatbot for E-commerce Websites. *Proceedings of ACL 2017, System Demonstrations*, 97–

102.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies–why and how. *Knowledge-Based Systems*, *6*(4), 258–266.

Dale, R. (2016). The return of the chatbots. *Natural Language Engineering*, *22*(5), 811–817.

Dale, R. (2020). Voice assistance in 2019. *Natural Language Engineering*, *26*(1), 129–136. https://doi.org/10.1017/S1351324919000640

De Maeyer, C., & Lee, M. (2022). I Feel You. *Human-Centered Software Engineering: 9th IFIP WG 13.2 International Working Conference, HCSE 2022, Eindhoven, The Netherlands, August 24–26, 2022, Proceedings*, 23–43.

De Maeyer, C., & Markopoulos, P. (2021). Experts' View on the Future Outlook on the Materialization, Expectations and Implementation of Digital Twins in Healthcare. *Interacting with Computers*, *33*(4), 380–394.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171--4186.

Diederich, S., Brendel, A. B., & Kolbe, L. M. (2019). *Towards a taxonomy of platforms for conversational agent design*.

Dinan, E., Abercrombie, G., Bergman, A. S., Spruit, S., Hovy, D., Boureau, Y.-L., & Rieser, V. (2021). Anticipating safety issues in e2e conversational ai: Framework and tooling. *ArXiv Preprint ArXiv:2107.03451*.

Duijst, D. (2017). *Can we Improve the User Experience of Chatbots with Personalisation* [Master's Thesis]. University of Amsterdam.

Dušek, O., Novikova, J., & Rieser, V. (2020). Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language*, *59*, 123–156.

Edu, J., Ferrer-Aran, X., Such, J., & Suarez-Tangil, G. (2022). Measuring Alexa Skill Privacy Practices across Three Years. *Proceedings of the ACM Web Conference 2022*, 670–680.

Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé III, H., Riener, A., & Riedl, M. O. (2022). Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–7. https://doi.org/10.1145/3491101.3503727

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A Taxonomy of Social Cues for Conversational Agents. *Int. J. Hum.-Comput. Stud.*, *132*, 138–161.

Fiore, S. M., Wiltshire, T. J., Lobato, E. J., Jentsch, F. G., Huang, W. H., & Axelrod, B. (2013). Toward understanding social cues and signals in human–robot interaction: Effects of robot gaze and proxemic behavior. *Frontiers in Psychology*, *4*, 859.

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young

adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Mental Health*, *4*(2), online.

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: Seven essential factors. *Science and Engineering Ethics*, *26*, 1771–1796.

Følstad, A., & Brandtzæg, P. B. (2017). Chatbots and the new world of HCI. *Interactions*, *24*(4), 38–42.

Følstad, A., Brandtzaeg, P. B., Feltwell, T., Law, E. L.-C., Tscheligi, M., & Luger, E. A. (2018). SIG: Chatbots for Social Good. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, SIG06:1-SIG06:4.

Følstad, A., & Skjuve, M. (2019). Chatbots for customer service: User experience and motivation. *Proceedings of the 1st International Conference on Conversational User Interfaces*, 1–9.

Frauenberger, C., Makhaeva, J., & Spiel, K. (2017). Blending methods: Developing participatory design sessions for autistic children. *Proceedings of the 2017 Conference on Interaction Design and Children*, 39–49.

Friedman, B. (1996). Value-sensitive design. *Interactions*, *3*(6), 16–23.

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. Mit Press.

Friedman, B., Kahn, P. H., Borning, A., & Huldtgren, A. (2013). Value sensitive design and information systems. *Early Engagement and New Technologies: Opening up the Laboratory*, 55–95.

Gabrielli, S., Rizzi, S., Carbone, S., Donisi, V., & others. (2020). A chatbot-based coaching intervention for adolescents to promote life skills: Pilot study. *JMIR Human Factors*, *7*(1), e16762.

Garg, R., Cui, H., Seligson, S., Zhang, B., Porcheron, M., Clark, L., Cowan, B. R., & Beneteau, E. (2022). The last decade of HCI research on children and voice-based conversational agents. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–19.

Garg, R., & Sengupta, S. (2019). " When you can do it, why can't I?": Racial and Socioeconomic Differences in Family Technology Use and Non-Use. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–22.

Garg, R., & Sengupta, S. (2020). Conversational technologies for in-home learning: Using co-design to understand children's and parents' perspectives. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Gentsch, P. (2019). Conversational AI: How (Chat) Bots Will Reshape the Digital Experience. In *AI in Marketing, Sales and Service* (pp. 81–125). Springer.

Glasser, A., Mande, V., & Huenerfauth, M. (2020). Accessibility for deaf and hard of hearing users: Sign language conversational user interfaces. *Proceedings of the 2nd Conference on Conversational User Interfaces*, 1–3.

Go, E., & Sundar, S. S. (2019). Humanizing Chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior*, *97*, 304–316.

Görnemann, E., & Spiekermann, S. (2022). Emotional responses to human values in technology: The case of conversational agents. *Human–Computer Interaction*, 1–28.

Graesser, A. C., Conley, M. W., & Olney, A. (2012). Intelligent tutoring systems. *APA Educational Psychology Handbook, Vol 3: Application to Learning and Teaching.*, 451–473.

Guerino, G. C., & Valentim, N. M. C. (2020). Usability and user experience evaluation of natural user interfaces: A systematic mapping study. *IET Software*, *14*(5), 451-467(16).

Guerino, G., Silva, W., Coleti, T., & Valentim, N. (2021). Assessing a Technology for Usability and User Experience Evaluation of Conversational Systems: An Exploratory Study. *Proceedings of the 23rd International Conference on Enterprise Information Systems - Volume 2: ICEIS,* 463–473. https://doi.org/10.5220/0010450204630473

Hampson, I., & Junor, A. (2005). Invisible work, invisible skills: Interactive customer service as articulation work. *New Technology, Work and Employment*, *20*(2), 166–181.

Han, S., & Lee, M. K. (2022). FAQ chatbot and inclusive learning in massive open online courses. *Computers & Education*, *179*, 104395.

Haugeland, I. K. F., Følstad, A., Taylor, C., & Bjørkli, C. (2022). Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design. *International Journal of Human-Computer Studies*, 102788.

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems*, *19*(2), 4.

Heyselaar, E., & Bosse, T. (2019). Using Theory of Mind to assess users' sense of agency in social chatbots. *Conversations 2019: 3rd International Workshop on Chatbot Research*, 1–13.

Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, *49*, 245–250.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hu, Y. (2019). *Do people want to message chatbots? Developing and comparing the usability of a conversational vs. menu-based chatbot in context of new hire onboarding.* [Master's Thesis, Aalto University]. https://aaltodoc.aalto.fi:443/handle/123456789/40834

Huang, M.-H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, *21*(2), 155–172.

Hwang, G.-J., & Chang, C.-Y. (2021). A review of opportunities and challenges of chatbots in education. *Interactive Learning Environments*, 1–14.

Inkster, B., Sarda, S., Subramanian, V., & others. (2018). An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR MHealth and UHealth*, *6*(11), e12106.

ISO 9241-11. (2018). *Ergonomics of human-system interaction: Part 11: Usability: Definitions and concepts*. International Organization for Standardization.

Jacobs, N. (2020). Capability sensitive design for health and wellbeing technologies. *Science and Engineering Ethics*, *26*(6), 3363–3391.

Jain, M., Kumar, P., Kota, R., & Patel, S. N. (2018). Evaluating and informing the design of chatbots. *Proceedings of the 2018 Designing Interactive Systems Conference*, 895–906.

Jakic, A., Wagner, M. O., & Meyer, A. (2017). The impact of language style accommodation during social media interactions on brand trust. *Journal of Service Management*, *28*(3), 418–441.

Jiang, R., & E Banchs, R. (2017). Towards Improving the Performance of Chat Oriented Dialogue System. *2017 International Conference on Asian Language Processing (IALP)*, 23–26.

Jonze, S. (Director). (2014). *Her*. Warner Bros. Picture presents an Annapurna Pictures production. https://search.library.wisc.edu/catalog/9910197336902121

Ju, W., & Leifer, L. (2008). The design of implicit interactions: Making interactive systems less obnoxious. *Design Issues*, *24*(3), 72–84.

Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Online. https://web.stanford.edu/~jurafsky/slp3/

Kaczorowska-Spychalska, D. (2019). How chatbots influence marketing. *Management*, *23*(1), 251–270.

Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J. T., & Bernstein, M. S. (2020). Conceptual metaphors impact perceptions of human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 1–26.

Khemani, K. H., & Reeves, S. (2022). Unpacking Practitioners’ Attitudes Towards Codifications of Design Knowledge for Voice User Interfaces. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–10. https://doi.org/10.1145/3491102.3517623

Khosrawi-Rad, B., Rinn, H., Schlimbach, R., Gebbing, P., Yang, X., Lattemann, C., Markgraf, D., & Robra-Bissantz, S. (2022). Conversational Agents in Education – A Systematic Literature Review. *ECIS 2022 Research Papers*. https://aisel.aisnet.org/ecis2022_rp/18

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will you accept an imperfect AI? exploring designs for adjusting end-user expectations of ai systems. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–14.

Kocielnik, R., Avrahami, D., Marlow, J., Lu, D., & Hsieh, G. (2018). Designing for workplace reflection: A chat and voice-based conversational agent. *Proceedings of the 2018 Designing Interactive Systems Conference*, 881–894.

Kraus, M., Schiller, M., Behnke, G., Bercher, P., Dorna, M., Dambier, M., Glimm, B., Biundo, S., & Minker, W. (2020). " Was that successful?" On Integrating Proactive Meta-Dialogue in a DIY-Assistant using Multimodal Cues. *Proceedings of the 2020 International Conference on Multimodal Interaction*, 585–594.

Kraus, M., Wagner, N., & Minker, W. (2021). Modelling and Predicting Trust for Developing Proactive Dialogue Strategies in Mixed-Initiative Interaction. *Proceedings of the 2021 International Conference on Multimodal Interaction*, 131–140.

Kwet, M. (2019). Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class*, *60*(4), 3–26.

Langevin, R., Lordon, R. J., Avrahami, T., Cowan, B. R., Hirsch, T., & Hsieh, G. (2021). Heuristic Evaluation of Conversational Agents. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3411764.3445312

Laumer, S., Maier, C., & Gubler, F. T. (2019). Chatbot acceptance in healthcare: Explaining user adoption of conversational agents for disease diagnosis. *Proceedings of the 27th European Conference on Information Systems (ECIS)*, 1–18.

Lee, H. R., Cheon, E., De Graaf, M., Alves-Oliveira, P., Zaga, C., & Young, J. (2019). Robots for social good: Exploring critical design for HRI. *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 681–682.

Lee, M., Ackermans, S., Van As, N., Chang, H., Lucas, E., & IJsselsteijn, W. (2019). Caring for Vincent: A chatbot for self-compassion. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Lee, M., & Contreras, J. (2023). Flourishing with Moral Emotions Through Conversational Agents. In M. Las Heras, M. Grau Grau, & Y. Rofcanin (Eds.), *Human Flourishing: A Multidisciplinary Perspective on Neuroscience, Health, Organizations and Arts* (pp. 163–179). Springer International Publishing. https://doi.org/10.1007/978-3-031-09786-7_11

Lee, M., Frank, L., & IJsselsteijn, W. (2021). Brokerbot: A cryptocurrency chatbot in the social-technical gap of trust. *Journal of Computer Supported Cooperative Work (JCSCW)*, *30*(1), 79–117.

Lee, M., Lucas, G., Mell, J., Johnson, E., & Gratch, J. (2019). What's on Your Virtual Mind?: Mind Perception in Human-Agent Negotiations. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 38–45. https://doi.org/10.1145/3308532.3329465

Lee, M., Noortman, R., Zaga, C., Starke, A., Huisman, G., & Andersen, K. (2021). Conversational

futures: Emancipating conversational interactions for futures worth wanting. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13.

Lee, M., Ruijten, P., Frank, L., de Kort, Y., & IJsselsteijn, W. (2021). People may punish, but not blame robots. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–11.

Lei, W., Zhang, Y., Song, F., Liang, H., Mao, J., Lv, J., Yang, Z., & Chua, T.-S. (2022). Interacting with Non-Cooperative User: A New Paradigm for Proactive Dialogue Policy. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 212–222.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.

Liao, S., Wilson, C., Cheng, L., Hu, H., & Deng, H. (2020). Measuring the effectiveness of privacy policies for voice assistant applications. *Annual Computer Security Applications Conference*, 856–869.

Liebrecht, C., Sander, L., & van Hooijdonk, C. (2020). Too Informal? How a Chatbot's Communication Style Affects Brand Attitude and Quality of Interaction. *Conversations 2020: 4th International Workshop on Chatbot Research*.

Lima, G., Grgić-Hlača, N., Jeong, J. K., & Cha, M. (2022). The Conflict Between Explainable and Accountable Decision-Making Algorithms. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2103–2113.

Ling, E. C., Tussyadiah, I., Tuomi, A., Stienmetz, J., & Ioannou, A. (2021). Factors influencing users' adoption and use of conversational agents: A systematic review. *Psychology & Marketing*, *38*(7), 1031–1051. https://doi.org/10.1002/mar.21491

Lister, K., Coughlan, T., Iniesto, F., Freear, N., & Devine, P. (2020). Accessible conversational user interfaces: Considerations for design. *Proceedings of the 17th International Web for All Conference*, 1–11.

Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., Jiang, Y., & Huang, M. (2021). Towards Emotional Support Dialog Systems. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3469–3483.

Llitjós, A. F. (2013). *IBM Design–A new Era at IBM. Lean UX leading theway*. https://submissions.agilealliance.org/system/attachments/attachments/000/000/306/original/IBM_Design_Thinking_Agile_2013.pdf

Lombard, M., & Xu, K. (2021). Social responses to media technologies in the 21st century: The media

are social actors paradigm. *Human-Machine Communication*, *2*, 29–55.

Lopatovska, I., Turpin, O., Davis, J., Connell, E., Denney, C., Fournier, H., Ravi, A., Yoon, J. H., & Parasnis, E. (2022). Capturing Teens' Voice in Designing Supportive Agents. *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–12.

Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94–100.

Lucas, G. M., Rizzo, A., Gratch, J., Scherer, S., Stratou, G., Boberg, J., & Morency, L.-P. (2017). Reporting mental health symptoms: Breaking down barriers to care with virtual human interviewers. *Frontiers in Robotics and AI*, *4*, 51.

Luger, E., & Sellen, A. (2016). Like Having a Really Bad PA: The Gulf between User Expectation and Experience of Conversational Agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297.

Madianou, M. (2021). Nonhuman humanitarianism: When "AI for good" can be harmful. *Information, Communication & Society*, *24*(6), 850–868.

Mahmood, T., Ricci, F., & Venturini, A. (2009). Improving recommendation effectiveness: Adapting a dialogue strategy in online travel planning. *Information Technology & Tourism*, *11*(4), 285–302.

Maslowski, I., Lagarde, D., & Clavel, C. (2017). In-the-wild chatbot corpus: From opinion analysis to interaction problem detection. *International Conference on Natural Language, Signal and Speech Processing*, 115–120.

Mateas, M. (1999). An Oz-Centric Review of Interactive Drama and Believable Agents. In M. J. Wooldridge & M. Veloso (Eds.), *Artificial Intelligence Today: Recent Trends and Developments* (pp. 297–328). Springer. https://doi.org/10.1007/3-540-48317-9_12

Mauldin, M. L. (1994). Chatterbots, Tinymuds, and the Turing Test Entering the Loebner Prize Competition. *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*, 16–21.

McTear, M. (2020). Conversational ai: Dialogue systems, conversational agents, and chatbots. *Synthesis Lectures on Human Language Technologies*, *13*(3), 1–251.

Meck, A.-M., Draxler, C., & Vogt, T. (2022). A Question of Fidelity: Comparing Different User Testing Methods for Evaluating In-Car Prompts. *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–5.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *International Conference on Learning Representations*.

Milne-Ives, M., Cock, C. de, Lim, E., Shehadeh, M. H., Pennington, N. de, Mole, G., Normando, E., & Meinert, E. (2020). The Effectiveness of Artificial Intelligence Conversational Agents in Health

Care: Systematic Review. *Journal of Medical Internet Research*, *22*(10), e20346. https://doi.org/10.2196/20346

Miner, A. S., Laranjo, L., & Kocaballi, A. B. (2020). Chatbots in the fight against the COVID-19 pandemic. *NPJ Digital Medicine*, *3*(1), 65.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.

Motger, Q., Franch, X., & Marco, J. (2022). Software-Based Dialogue Systems: Survey, Taxonomy, and Challenges. *ACM Computing Surveys*, *55*(5), 91:1-91:42. https://doi.org/10.1145/3527450

Muller, M., & Liao, Q. V. (2017). Exploring AI Ethics and Values through Participatory Design Fictions. *Human Computer Interaction Consortium*.

Murad, C., Tasnim, H., & Munteanu, C. (2022). Voice-First Interfaces in a GUI-First Design World: Barriers and Opportunities to Supporting VUI Designers On-the-Job. *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–10. https://doi.org/10.1145/3543829.3543842

Nardi, B. A., & ENgeström, Y. (1999). A Web on the Wind: The Structure of Invisible Work. *Computer Supported Cooperative Work*, *8*(1–2), 1–8. https://doi.org/10.1023/A:1008694621289

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78.

Nawaz, A. (2012). A comparison of card-sorting analysis methods. *10th Asia Pacific Conference on Computer Human Interaction (Apchi 2012). Matsue-City, Shimane, Japan*, 28–31.

Neururer, M., Schlögl, S., Brinkschulte, L., & Groth, A. (2018). Perceptions on Authenticity in Chat Bots. *Multimodal Technologies and Interaction*, *2*(3), 60.

Nguyen, Q. N., Sidorova, A., & Torres, R. (2021). User interactions with chatbot interfaces vs. Menu-based interfaces: An empirical study. *Computers in Human Behavior*, 107093.

Nothdurft, F., Behnke, G., Bercher, P., Biundo, S., & Minker, W. (2015). The interplay of user-centered dialog systems and AI planning. *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 344–353.

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and Engineering Ethics*, *24*(4), 1201–1219.

Pearson, J., Robinson, S., Reitmaier, T., Jones, M., Ahire, S., Joshi, A., Sahoo, D., Maravi, N., & Bhikne, B. (2019). StreetWise: Smart speakers vs human help in public slum settings. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–13.

Pelachaud, C. (2017). Greta: A conversing socio-emotional agent. *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*, 9–10.

Perski, O., Crane, D., Beard, E., & Brown, J. (2019). Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. *Digital Health*, *5*, 2055207619880676.

Pillai, R., & Sivathanu, B. (2020). Adoption of AI-based chatbots for hospitality and tourism. *International Journal of Contemporary Hospitality Management*.

Polaine, A., Løvlie, L., & Reason, B. (2013). *Service design: From insight to implementation*. Rosenfeld media.

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–12.

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, *37*, 98–125.

Portela, M., & Granell-Canut, C. (2017). A new friend in our Smartphone? Observing Interactions with Chatbots in the search of emotional engagement. *Interaccion*, 48:1-48:7.

Preum, S. M., Munir, S., Ma, M., Yasar, M. S., Stone, D. J., Williams, R., Alemzadeh, H., & Stankovic, J. A. (2021). A review of cognitive assistants for healthcare: Trends, prospects, and future directions. *ACM Computing Surveys (CSUR)*, *53*(6), 1–37.

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 7:1-7:23. https://doi.org/10.1145/3449081

Rashita, R., Himanshu, J., & Vineet, K. (2021). *Conversational AI Market Size, Share & Growth | Trends—2030* (No. A13682; SE: Emerging and Next Generation Technologies, p. 288). Allied Market Research. https://www.alliedmarketresearch.com/conversational-ai-market-A13682

Rattenbury, T., Hellerstein, J. M., Heer, J., Kandel, S., & Carreras, C. (2017). *Principles of data wrangling: Practical techniques for data preparation*. O'Reilly Media, Inc.

Reeves, B., & Nass, C. (1996). *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press.

Ren, R., Castro, J. W., Acuña, S. T., & de Lara, J. (2019). Usability of chatbots: A systematic mapping study. *Proc. 31st Int. Conf. Software Engineering and Knowledge Engineering*, 479–484.

Rick, S. R., Goldberg, A. P., & Weibel, N. (2019). SleepBot: Encouraging sleep hygiene using an intelligent chatbot. *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*, 107–108.

Rieser, V., Lemon, O., & Keizer, S. (2014). Natural language generation as incremental planning under uncertainty: Adaptive information presentation for statistical dialogue systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(5), 979–994.

Ringfort-Felner, R., Laschke, M., Sadeghian, S., & Hassenzahl, M. (2022). Kiro: A Design Fiction to Explore Social Conversation with Voice Assistants. *Proceedings of the ACM on Human-Computer Interaction*, *6*(GROUP), 1–21.

Robeyns, I. (2009). Capability approach. In *Handbook of economics and ethics*. Edward Elgar Publishing.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., & others. (2021). Recipes for Building an Open-Domain Chatbot. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 300–325.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1).

Samson, B. P. V., & Sumi, Y. (2020). Are Two Heads Better than One? Exploring Two-Party Conversations for Car Navigation Voice Guidance. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3334480.3382818

Schaffer, J., Playa Vista, C., O'Donovan, J., Michaelis, J., Adelphi, M., Raglin, A., & Höllerer, T. (2019). I Can Do Better Than Your AI: Expertise and Explanations. *Proceedings of the 2019 ACM International Conference on Intelligent User Interfaces (IUI'19)*.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*(2), 361–382.

Schlesinger, A., O'Hara, K. P., & Taylor, A. S. (2018). Let's Talk About Race: Identity, Chatbots, and AI. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 315:1-315:14.

Schmidt, K., & Schmidt, K. (2011). Remarks on the complexity of cooperative work (2002). *Cooperative Work and Coordinative Practices: Contributions to the Conceptual Foundations of Computer-Supported Cooperative Work (CSCW)*, 167–199.

Schneider, J., & Stickdorn, M. (2011). *This is service design thinking: Basics, tools, cases*. Wiley.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Seering, J., Luria, M., Ye, C., Kaufman, G., & Hammer, J. (2020). It takes a village: Integrating an adaptive chatbot into an online gaming community. *Proceedings of the 2020 Chi Conference on Human Factors in Computing Systems*, 1–13.

Seidelin, C., Dittrich, Y., & Grönvall, E. (2020). Foregrounding data in co-design – An exploration of how data may become an object of design. *International Journal of Human-Computer Studies*, *143*, 102505. https://doi.org/10.1016/j.ijhcs.2020.102505

Sharma, S., Brennan, J., & Nurse, J. (2021). StockBabble: A conversational financial agent to support

stock market investors. *Proceedings of the 3rd Conference on Conversational User Interfaces*, 1–5.

Shechtman, N., & Horowitz, L. M. (2003). Media inequality in conversation: How people behave differently when interacting with computers and people. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 281–288.

Shorter, M., Minder, B., Rogers, J., Baldauf, M., Todisco, A., Junginger, S., Aytaç, A., & Wolf, P. (2022). Materialising the Immaterial: Provotyping to Explore Voice Assistant Complexities. *Designing Interactive Systems Conference*, 1512–1524. https://doi.org/10.1145/3532106.3533519

Simonsen, L., Steinstø, T., Verne, G., & Bratteteig, T. (2020). "I'm disabled and married to a foreign single mother". Public service chatbot's advice on citizens' complex lives. *Electronic Participation: 12th IFIP WG 8.5 International Conference, EPart 2020, Linköping, Sweden, August 31–September 2, 2020, Proceedings 12*, 133–146.

Simpson, J., Stening, H., Nalepka, P., Dras, M., Reichle, E. D., Hosking, S., Best, C. J., Richards, D., & Richardson, M. J. (2022). DesertWoZ: A Wizard of Oz Environment to Support the Design of Collaborative Conversational Agents. *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing*, 188–192.

Søndergaard, M. L. J., & Hansen, L. K. (2018). Intimate futures: Staying with the trouble of digital personal assistants through design fiction. *Proceedings of the 2018 Designing Interactive Systems Conference*, 869–880.

Spiel, K., Gerling, K., Bennett, C. L., Brulé, E., Williams, R. M., Rode, J., & Mankoff, J. (2020). Nothing about us without us: Investigating the role of critical disability studies in HCI. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–8.

Stoilova, E. (2021). AI chatbots as a customer service and support tool. *ROBONOMICS: The Journal of the Automated Economy*, *2*, 21–21.

Storey, M.-A., & Zagalsky, A. (2016). Disrupting developer productivity one bot at a time. *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, 928–931.

Strengers, Y., & Kennedy, J. (2021). *The smart wife: Why Siri, Alexa, and other smart home devices need a feminist reboot*. Mit Press.

Strengers, Y., Qu, L., Xu, Q., & Knibbe, J. (2020). Adhering, steering, and queering: Treatment of gender in natural language generation. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Suchman, L., & Suchman, L. A. (2007). *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.

Tanaka, F., & Matsuzoe, S. (2012). Children teach a care-receiving robot to promote their learning:

Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction*, *1*(1), 78–95.

Thomaz, F., Salge, C., Karahanna, E., & Hulland, J. (2020). Learning from the Dark Web: Leveraging conversational agents in the era of hyper-privacy to enhance marketing. *Journal of the Academy of Marketing Science*, *48*(1), 43–63.

Traum, D. R., & Hinkelman, E. A. (1992). Conversation acts in task-oriented spoken dialogue. *Computational Intelligence*, *8*(3), 575–599.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B. W., & Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE Journal of Selected Topics in Signal Processing*, *11*(8), 1301–1309.

Ueda, K., & Yamada, Y. (2023). *ChatGPT is not an author, but then, who is eligible for authorship?*

Ukpabi, D. C., Aslam, B., & Karjaluoto, H. (2019). Chatbot Adoption in Tourism Services: A Conceptual Exploration. In S. Ivanov & C. Webster (Eds.), *Robots, Artificial Intelligence, and Service Automation in Travel, Tourism and Hospitality* (pp. 105–121). Emerald Publishing Limited. https://doi.org/10.1108/978-1-78756-687-320191006

Ultes, S., Kraus, M., Schmitt, A., & Minker, W. (2015). Quality-adaptive spoken dialogue initiative selection and implications on reward modelling. *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 374–383.

Valério, F. A., Guimarães, T. G., Prates, R. O., & Candello, H. (2020). Comparing users' perception of different chatbot interaction paradigms: A case study. *Proceedings of the 19th Brazilian Symposium on Human Factors in Computing Systems*, 1–10.

van der Goot, M. J., Hafkamp, L., & Dankfort, Z. (2020). Customer service chatbots: A qualitative interview study into the communication journey of customers. *Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event, November 23–24, 2020, Revised Selected Papers 4*, 190–204.

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, *614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7

van Hooijdonk, C., Martijn, G., & Liebrecht, C. (2023). A Framework and Content Analysis of Social Cues in the Introductions of Customer Service Chatbots. *Chatbot Research and Design: 6th International Workshop, CONVERSATIONS 2022, Amsterdam, The Netherlands, November 22–23, 2022, Revised Selected Papers*, 118–133.

Vapnik, V. (2000). SVM method of estimating density, conditional probability, and conditional density. *2000 IEEE International Symposium on Circuits and Systems (ISCAS)*, *2*, 749–752.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \Lukasz, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Venkatesh, V., Thong, J. Y., Chan, F. K., Hu, P. J.-H., & Brown, S. A. (2011). Extending the two-stage information systems continuance model: Incorporating UTAUT predictors and the role of context. *Information Systems Journal*, *21*(6), 527–555.

Wallace, R. S. (2009). The anatomy of A.L.I.C.E. In *Parsing the Turing Test* (pp. 181–210). Springer Netherlands.

Wallis, P., & Norling, E. (2005). The Trouble with Chatbots: Social skills in a social world. *Proceedings of the Joint Symposium on Virtual Social Agents*, 29–38.

Wambsganss, T., Höch, A., Zierau, N., & Söllner, M. (2021). Ethical design of conversational agents: Towards principles for a value-sensitive design. *Innovation Through Information Systems: Volume I: A Collection of Latest Research on Domain Issues*, 539–557.

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

Wessel, M., de Souza, B. M., Steinmacher, I., Wiese, I. S., Polato, I., Chaves, A. P., & Gerosa, M. A. (2018). The Power of Bots: Characterizing and Understanding Bots in OSS Projects. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW). https://doi.org/10.1145/3274451

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

Williams, J. D., Raux, A., & Henderson, M. (2016). The dialog state tracking challenge series: A review. *Dialogue & Discourse*, *7*(3), 4–33.

Wiltshire, T. J., Snow, S. L., Lobato, E. J., & Fiore, S. M. (2014). Leveraging social judgment theory to examine the relationship between social cues and signals in human-robot interactions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 1336–1340.

Winkler, R., Hobert, S., Salovaara, A., Söllner, M., & Leimeister, J. M. (2020). Sara, the lecturer: Improving learning in online education with a scaffolding-based conversational agent. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14.

Winograd, T. (1971). *Procedures as a representation for data in a computer program for understanding natural language*. DTIC Document.

Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why we should have seen that coming: Comments on Microsoft's tay" experiment," and wider implications. *Acm Sigcas Computers and Society*, *47*(3), 54–64.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there

yet?-A systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, *4*, 654924.

Wu, Y., Porcheron, M., Doyle, P., Edwards, J., Rough, D., Cooney, O., Bleakley, A., Clark, L., & Cowan, B. (2022). Comparing Command Construction in Native and Non-Native Speaker IPA Interaction through Conversation Analysis. *Proceedings of the 4th Conference on Conversational User Interfaces*, 1–12.

Wube, H. D., Esubalew, S. Z., Weldesellasie, F. F., & Debelee, T. G. (2022). Text-based chatbot in financial sector: A systematic literature review. *Data Science in Finance and Economics*, *2*(3), 232–259.

Xuetao, M., Bouchet, F., & Sansonnet, J.-P. (2009). Impact of agent's answers variability on its believability and human-likeness and consequent chatbot improvements. *Proc. of AISB*, 31–36.

Yanishevskaya, N., Kuznetsova, L., Lokhacheva, K., Zabrodina, L., Parfenov, D., & Bolodurina, I. (2019). Application of Intelligent Algorithms for the Development of a Virtual Automated Planning Assistant for the Optimal Tourist Travel Route. *International Conference of Artificial Intelligence, Medical Engineering, Education*, 13–22.

Youn, S., & Jin, S. V. (2021). In AI We Trust?" The Effects of Parasocial Interaction and Technopian versus Luddite Ideological Views on Chatbot-Based Customer Relationship Management in the Emerging "Feeling Economy. *Computers in Human Behavior*, 106721.

Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., & Yu, K. (2010). The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, *24*(2), 150–174.

Young, S., Gašić, M., Thomson, B., & Williams, J. D. (2013). Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, *101*(5), 1160–1179.

Yu, D., & Deng, L. (2016). *Automatic speech recognition* (Vol. 1). Springer.

Zaga, C., Lohse, M., Truong, K. P., & Evers, V. (2015). The effect of a robot's social character on children's task engagement: Peer versus tutor. *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*, 704–713.

Zembylas, M. (2023). A decolonial approach to AI in higher education teaching and learning: Strategies for undoing the ethics of digital neocolonialism. *Learning, Media and Technology*, *48*(1), 25–37.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-tuning language models from human preferences. *ArXiv Preprint ArXiv:1909.08593*.

Zue, V. W., & Glass, J. R. (2000). Conversational interfaces: Advances and challenges. *Proceedings of*

the IEEE, *88*(8), 1166–1180.