All Work and No Play? Conversations with a Question-and-Answer Chatbot in the Wild

Q. Vera Liao[1], **Muhammed Mas-ud Hussain**[4 *], **Praveen Chandar**[6 †], **Matthew Davis**[2], **Yasaman Khazaen**[2], **Marco Patricio Crasso**[3], **Dakuo Wang**[1], **Michael Muller**[2], **N. Sadat Shami**[5], **Werner Geyer**[2]
IBM Research, [1]Yorktown Heights, NY, USA; [2]Cambridge, MA, USA; [3]Capital Federal, Argentina
[4]Northwestern University, Evanston, IL, USA
[5]IBM, Armonk, NY, USA
[6]Spotify Research, New York, NY, USA
{vera.liao, dakuo.wang}@ibm.com, {davismat, yasaman.khazaeni, michael_muller, sadat, werner.geyer}@us.ibm.com, mas-ud@u.northwestern.edu, praveenr@spotify.com, crasso@ar.ibm.com

ABSTRACT
Many conversational agents (CAs) are developed to answer users' questions in a specialized domain. In everyday use of CAs, user experience may extend beyond satisfying information needs to the enjoyment of conversations with CAs, some of which represent *playful* interactions. By studying a field deployment of a Human Resource chatbot, we report on users' interest areas in conversational interactions to inform the development of CAs. Through the lens of statistical modeling, we also highlight rich signals in conversational interactions for inferring user satisfaction with the instrumental usage and playful interactions with the agent. These signals can be utilized to develop agents that adapt functionality and interaction styles. By contrasting these signals, we shed light on the varying functions of conversational interactions. We discuss design implications for CAs, and directions for developing adaptive agents based on users' conversational behaviors.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords
Conversational agent; chatbot; dialog system; human-agent interaction; playful; adaption; user modeling

INTRODUCTION
There is a growing excitement around conversational agents (CAs) or "chatbots". From tech giants' core products such as Apple Siri, Amazon Alexa, IBM Watson, to numerous startup companies, many are compelled by the idea of advances in artificial intelligence combined with a natural form of interactions. However, before this wave of marketing hype, research on CAs has come a long way in the past half century, but also saw several unfortunate failures in public reception (e.g. [**?**, **?**]). Two points of criticism have been frequently raised for studies of CAs. One is a lack of understanding on real-life user experience and attention to the gap between user interactions *in the lab* and those *in the wild* [**?**, **?**]. Another point is the focus on narrowly constrained agent initiated conversations for the task domains, which provides little information about user interests in conversing with CAs for future system development [**?**]. Although many recent popular CAs, often in the form of an intelligent personal assistant, provide free-form text input interfaces that invite users to "ask me anything", there is surprisingly limited empirical account of how users converse with these agents in the wild. This poses a challenge as the development of CAs, at least in the near term, relies heavily on the *anticipation* of what users may say to the agent.

To fill these gaps, we study a field deployment of a question-and-answer (QA) CA. Specifically, a Human Resource (HR) chatbot provided company-related information assistance to 377 new employees for 6 weeks. Although the CA functions as a QA system, the focus of this paper is on users' *conversational interactions*, or social dialogues, with the agent (36% of interaction logs). Such interactions are often abundant as CAs naturally elicit social behaviors with a human role metaphor. Meanwhile, there is a tradition of separating communicative and task-oriented interactions in developing CAs [**?**, **?**], and considering the former to be more unbounded and thus challenging to anticipate, but have the advantage of higher generalizability across domains [**?**]. The generalizability motivated some to build domain-independent conversation architectures to accelerate the development of CAs [**?**].

Also underscoring the necessity of studying conversational interactions is the rich signals they may carry for inferring user status. This has been a longstanding interest in the related embodied conversational agents (ECA) and human-robot interaction (HRI) communities [**?**, **?**], as foundational work to build

* This work was done during an internship at IBM
† This work was done while working at IBM

CHI 2018, April 21–26, 2018, Montreal, QC, Canada
 ISBN 978-1-4503-5620-6/18/04…$15.00
DOI: https://doi.org/10.1145/3173574.3173577

adaptive agents that can attend to user needs. For example, by inferring a decline in user engagement, an agent can immediately employ strategies to re-engage the user [**?**]. To infer such an internal user status, agents rely on recognizing *signals* in users' behavioral manifestation. For example, gaze fixation [**?**] and attentive feedback ("*un-huh*") [**?**] are signals of engagement. These association rules are an integral part of the computational models underlying adaptive agents. However, most existing work drew on observations from human-human communications, and aimed to infer human concepts of interpersonal status such as rapport [**?**, **?**] and trust [**?**]. In the context of un-embodied QA agents, the system is not intended to serve full conversations, and some users may simply reject to anthropomorphize QA agents such as Siri [**?**]. It is arguable whether these behavioral signals known from human conversations still hold. It is also arguable whether human-like inter-personal status should be the primary dimension that a QA agent is concerned with.

In this work, we explore the associations between signals in conversational interactions and user satisfaction with the QA agent. Through the lens of statistical modeling, we take an empirical, data-driven approach to inform areas to obtain such signals, and their potential deviation from human conversations. We ask questions such as whether it is a reliable signal for user satisfaction, if a user praises the agent by saying "*you are smart*"; or whether it signals a real user frustration that the system should attend to, if a user tells the agent to "*shut up*." For a QA system, knowing these kinds of signals can enable real-time adaption of algorithms and other system functions.

Meanwhile, unlike conventional information systems, user experience with CAs may extend beyond the instrumental usage. As seen in the interaction logs of our agent, a large portion were human-like conversations unrelated to the QA functions. Recent studies considered this kind of behaviors as *playful* interactions and a key aspect of the adoption of CAs [**?**, **?**, **?**], through which users explore the system and seek satisfaction from a sense of social contact. Studies also reported that such a tendency varies between individuals, and may reflect a fundamental difference in the orientation of attributing lifelike qualities to a CA versus purely instrumental values. Such an orientation may lead to differences in how users evaluate a CA. While playful users seek satisfaction from agents' "humanized or humorous responses" [**?**], those taking an instrumental view have a very different set of system preferences [**?**]. But how can an agent recognize signals for such an individual difference and adapt its interaction styles?

Motivated by these questions, we study conversational interactions form the log data of the field deployment. With self-reported satisfaction from survey responses, we employ a penalized regression analysis to identify associations between conversational signals and user satisfaction with the agent's instrumental usage and playful interactions. We ask:

- **RQ1:** What kinds of conversational interactions did users have with the QA agent in the wild?
- **RQ2:** What kinds of conversational interactions can be used as signals for inferring user satisfaction with the agent's functional performance, and playful interactions?

Research contributions from this work are threefold: 1) the results characterize users' conversational interactions with a QA agent in the wild; 2) the results suggest rich signals in conversational interactions for inferring user satisfaction status, which can be utilized to develop adaptive agents; 3) by contrasting the signals for functionality and playfulness, the results provide nuanced understanding of the underlying functions of users' conversational behaviors. For example, while some conversations are carried out with evident playful intentions, others may serve primarily instrumental purposes.

BACKGROUND AND RELATED WORK

We first discuss prior work on CAs and our focus on conversational interactions. Drawing on research on system adaption in relevant fields, we consider the potentials of utilizing signals in conversational interactions for inferring user satisfaction. Lastly, we motivate our focus on functionality and playfulness by discussing prior work on user experience with CAs.

Conversational interactions with CAs

Development of CAs can be dated back to the 1950s with prominent examples such as ELIZA [**?**]. Within the HCI field, research largely focused on embodied CAs. Anthropomorphism is emphasized in multiple modalities to regulate human-computer interactions in a familiar way and to manifest social intelligence such as trustworthiness [**?**, **?**]. Recently, the term "chatbot" is used to refer to CAs that employ primarily text-based or speech-based input without embodied modalities. This type of CA has become mainstream products. Some argue that for these systems, anthropomorphism is no longer a principal goal [**?**], and the single modality directs more attention to task performances [**?**], especially since many of these CAs are core components of utility applications. Despite the anti-anthropomorphism argument, the interaction is still based on the metaphor of human conversations, which is a complex machinery in its own right [**?**], but can also diverge from human conversations in many ways [**?**].

This highlights the necessity of studying patterns of conversational interactions with CAs, which can be considered as *user utterances in performing communicative and social functions instead of task-oriented functions (e.g.,QA query)*. Early systems often adopted agent controlled conversations to avoid the daunting challenge of handling unbounded conversations initiated by users (e.g. [**?**]). But this approach is inadequate for realistic conversational capabilities, and obsolete for QA agents that provide information assistance through free-form text input. At the present time, whether using a rule-based system or advanced technologies such as discourse planners, the development of CAs relies heavily on the anticipation of what users may say to the agent. Ignoring common patterns may result in the absence of necessary system knowledge, and thus repeatedly frustrating "*sorry I didn't get it*" responses. To overcome the problem, development of CAs has to follow a laborious iterative process to bootstrap from user data [**?**].

The hope lies in the fact that there are tractable, domain-independent patterns in conversations. A longstanding interest in linguistics research is to develop schemas of conversational acts to describe the performative functions of utterances in a content-independent way (e.g. [**?**, **?**]). Although these schemas informed the development of many CAs, there is fairly limited

shared knowledge on what are the common categories of user conversations with CAs. One exception is Kopp et al. [**?**], which described how people conversed with Max, a museum guide agent working in a real-world setting, by summarizing eight categories of user utterances. [**?**, **?**] are two other studies providing similar characterizations, but all reported on speech-based embodied CAs. Despite the popularity of commercial chatbots (e.g. Siri), there is little empirical report of user interaction patterns except for a few qualitative studies [**?**, **?**]. Our work is motivated to fill this gap by providing an empirical account of users' interest areas in conversations with text-based QA agents.

Inferring user status from conversational behaviors

Our focus on conversational interactions is also motivated by the potentially rich signals in them to infer user status, as a first step towards building adaptive agents. Dynamic adaption is arguably one of the most important values that CAs can deliver with a human based metaphor [**?**, **?**, **?**]. In early systems, inferring user status relied largely on manually specified rules by drawing on theories or observations from human communications (e.g., nodding means positive feedback). Recent work explored data-driven methods to establish associations between behavioral signals and user status in a principled fashion. The most notable work is by Cassell et al. on modeling behavioral signals for social status between interlocutors, as part of the research towards "socially aware CAs". This line of work used labeled social status such as rapport and politeness in episodes of human tutor videos as ground truth to build predictive models with verbal and non-verbal behaviors. It provided nuanced insights on the associations between social status and manifested behaviors. Despite the fruitful results, they may have limited applicability for QA chatbots. Compared to embodied tutoring CAs, user behaviors with QA chatbots may share far less similarities with human conversations. Moreover, unlike tutoring activities that require continuous engagement, social status may not be the dimension of primary concern for QA agents.

By considering QA agents as information-retrieval (IR) systems with a conversational interface [**?**], we resort to the volume of IR literature on adaption, where a primary focus is on adapting algorithms for less satisfied users who may have different information needs or expectations. Because obtaining explicit satisfaction is costly or infeasible at times, an area that draws continuous interest is to infer user satisfaction from behavioral signals such as dwell time and word choices as "implicit feedback" (see review in [**?**]). By monitoring implicit feedback, real-time algorithmic adaption is a highly desirable possibility. Research also explored adapting other functions, e.g. providing query assistance to unsatisfied users [**?**]. A common approach to identify implicit feedback is to study what behaviors predict satisfaction, often gathered by surveys. To develop automatic evaluation of chatbots, [**?**] collected user satisfaction with Microsoft Cortana performance through a survey and modeled predictive actions (e.g. follow-up inquiries signal dissatisfaction). While [**?**] focused on QA queries, we aim to highlight the rich opportunities in conversational interactions to obtain signals for user satisfaction, which would in itself advocate the use of conversational interfaces.

Conversations with CAs: beyond functionality

We also look beyond satisfaction with task performances. Studies evaluating CAs [**?**, **?**, **?**] repeatedly found user satisfaction strongly impacted by social designs such as the agent's representation [**?**] and personality [**?**]. According to Justine Cassell, CAs should target similar goals of human conversations, which are to fulfill *propositional goals*—conveying meaningful information, and *interactional goals*—ensuring the communication process to be enjoyable [**?**, **?**]. However, the interactional goal with CAs may differ from that of human. [**?**] conducted a lab study comparing conversations with a human versus an agent. They found that behaviors associated with relationship building (e.g., sharing opinions) to happen much less with the agent. This means, expectedly, users had less conscious relational motivation in interactions with the agent. In contrast, recent studies of agents in the wild reported rich relational behaviors, both positive and negative ones. This highlights that certain interactions with CAs may only be observed in the wild. [**?**] studied how people talked to a receptionist CA and identified three types of relational behaviors—politeness (e.g., greeting), sociable behaviors (e.g., small talk), and negative behaviors (e,g., insult). In studying Max, [**?**] reported that more than one-third of questions were small talks. Meanwhile, they observed patterns that are rare in human conversations, including flaming behaviors and "testing of intelligence" with unrealistic questions.

Although on the surface these utterances resemble chit-chat in human communications, they do not necessarily indicate conscious compliance of social norms, meaning that they may not signal the same underlying intentions as in human conversations. Similarly, negative comments may not signal actual negativity towards the agent. In a recent study on everyday use of CA, Luger and Sellen [**?**] proposed the notion of "playing as a point of entry" to the adoption of CAs by *asking random questions and looking for humanized or humorous responses*, as a way to explore the system and satisfy the desire for entertainment. Other studies showed that students engaged in playful interactions such as making face-threatening comments with tutoring agents, and found them to improve learning experience [**?**]. According to the theory of anthropomorphism [**?**], these behaviors are not only to satisfy sociality needs through a sense of social contact, but also to reduce uncertainty about the system to interact more effectively.

Playfulness is not a foreign concept to the HCI community. In the 1980s, with the introduction of personal computers in organizations, researchers noted its suitability for inviting a sense of playfulness— *a tendency to interact spontaneously and imaginatively with computer systems* [**?**, **?**]. Webster et al. considered playfulness as an individual and situation specific feature that predicts technology usage patterns [**?**], and suggested design guidelines to promote playfulness [**?**]. Recently, [**?**] proposed a Technology Acceptance Model (TAM) for interface agent of email systems and added playfulness to the original TAM, suggesting that "*individuals may explore playful features of the agent and find it highly enjoyable*."

We note that in these studies the definition of playfulness point to two intertwined aspects—one is deliberately anthropomor-

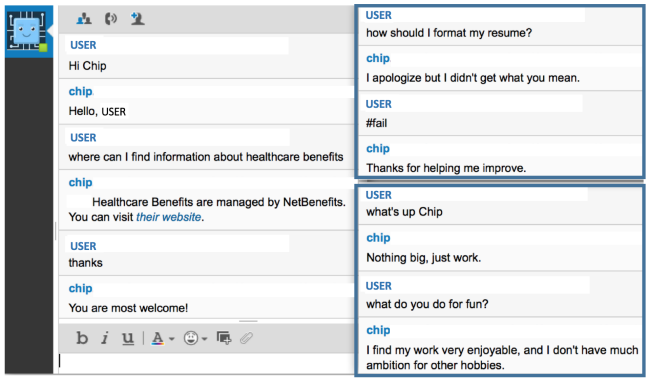

Figure 1. Left: user asked a work-related question with opening and acknowledging the answer; Top right: user asked a question that Chip could not understand then gave *#fail*; Bottom right: playful interactions

phizing the agent, the other is actively conversing about topics outside the functional scope. Previous work also studied *individual differences* in these kinds of behaviors under the terms *agent sociality* [**?**] and *social schema* [**?**], and argued that they reflect an orientation of users' mental schema of a CA as a sociable entity versus a purely instrumental tool [**?**, **?**, **?**]. In our previous work [**?**], using a self-reported measure on agent sociality orientation—tendency to engage in social conversations with CAs, we interviewed users varying on the orientation and found remarkable differences in their preference for interaction styles of CAs. For example, users viewing a CA as an instrumental tool prefer agent responses resembling search engine output, consider humanized features to be unnecessary, and may be turned away by lengthy conversations, while those with high sociality orientation desire for natural conversations and the agent to show more personalities.

Following the prior work, we argue that playfulness can be a key interactional goal for the usage of CAs. We use the term "*playfulness*" to refer to seeking satisfaction from off-task human-like conversations with a QA agent based on the above definition, without presuming that they carry the same meaning as in human communications. While the prior work motivates *why* designing CAs for playful interactions, and suggests *how* to design for playfulness, our goal is to inform *for whom* to design for playfulness by identifying behavioral signals of users seeking satisfaction from playful interactions through a data-driven approach. The results would move beyond knowledge from previous studies showing correlations between playfulness or sociality orientation with selected conversational markers such as greeting the agent [**?**], to form more comprehensive understanding on what kinds of user behaviors are carried out with playful intentions.

SYSTEM DESCRIPTION

A CA called Cognitive Human Interface Personality, or *Chip* for short was developed to provide *company related information assistance*. Targeted users of this deployment were new hires of a large multi-national enterprise, who had frequent HR information needs to orient around the company and complete administrative tasks. Chip is installed on a company-wide Instant Messenger (IM) tool to send and receive messages. Figure 1 presents multiple examples of conversations a user had with Chip, including both QAs and playful chit-chat.

The main function of Chip is to answer company-related questions such as "*tell me about my health benefits*" or "*how can I find IT help*". Chip's answers are mostly in the form of curated texts, some containing links to web pages with more detailed information. On occasions where Chip could not understand the user input, it says "*Sorry, I could not understand your question*" or variations thereof. Chip can also perform search tasks by accessing other applications. For example, when receiving a question it could not answer, it accesses an internal search engine and outputs snippets it returns, if available. When a user asks to "*look up [name] [phone number/location/...]*" or "*look for experts in ...*", Chip can retrieve such information from other internal applications. As with any CA, Chip may give wrong answers. We suggested to participants that they could respond "*#fail*" to give feedback, which would help improve Chip in the future.

We designed Chip with an HR assistant persona and made handling common conversational interactions a design goal from early on. To anticipate non-QA input is a nontrivial task. We resorted to two approaches. First, we followed an iterative design process by conducting multiple pilot studies ranging from 10 to 30 users, and bootstrapped the development of conversations from the data. Meanwhile, we referred to a package in IBM Watson Dialog[1] that provides instances of chit-chat collected from previous deployment of CAs. Another way we attempted to make Chip more social is to have it pro-actively send reminder messages. Some of them were to remind users of the availability of functions (e.g."*I can help find information about your colleagues...*"). Others were about tasks that new hires had to complete, such as filling out forms. The reminders were sent twice per week, and were triggered when the users logged in the IM tool on the scheduled day.

Natural language classifiers (NLCs) and performance

For the analysis, we utilized the natural language classifiers (NLCs) of the system to provide a characterization of users' conversational input. Here we briefly discuss the NLCs to provide background for our methodology. The technical details of the NLCs are beyond the scope of this paper. Many current CA technologies rely on NLCs to classify a user input (e.g., "*hello*") into a higher-level category of *intent* (e.g., "GREETING") known to the system in order to retrieve answers. Chip adopted a multi-level NLC model by *independently* training two levels of NLCs [**?**], each as a multi-class classifier (i.e., an input is classified to be the intent class with the highest confidence score). NLC1 contains higher-level categories of intent, each has several matching NLC2 sub-categories. For example, when a user asks "*tell me about health benefits*", it will be classified as "BENEFITS" by NLC1, and "health benefits" by NLC2, independently. "Health benefits" is a sub-intent known to match "BENEFITS", which also has other matching sub-intents such as "dental plan" and "employee discount". Each NLC2 class is linked to a curated answer, sometimes with variations to be randomly retrieved. For example, in this case, Chip will output the curated answer linked to "health benefits" to answer the user question. With this setup, user input fell into three categories:

[1] https://www.ibm.com/watson/developercloud/dialog.html

	I-correct: correct intent	*I-incorrect: incorrect intent*	*I-low: low confidence*
Percentage	74.7%	6.3%	19.0%
P (evaluated positive)	87.1%	25%	-
P (conversational)	80.6%	13.1%	6.3%

Table 1. Information of the three categories of user input

- *Correct intent recognition (I-correct)*: When classified NLC2 is a matching sub-category of NLC1, we expect Chip to have given mostly reasonable answers.
- *Incorrect intent recognition (I-incorrect)*: When classified NLC2 is not a matching sub-category of NLC1, often due to questions not anticipated thus no training examples were given during the development [2], we expect Chip to have given low-quality answers.
- *Low confidence (I-low)*: when either NLC was below a confidence threshold, Chip replied "*Sorry I didn't understand*".

Table 1 shows the distributions of user input in the three categories. Two researchers did a binary evaluation of the answer quality (reasonable/unreasonable) with 140 input-output pairs randomly drawn from *I-correct* category and 40 pairs from *I-incorrect* (Cohen's $\kappa = 0.84$). As expected, more than 87% of user input in *I-correct* received reasonable answers, and only 25% for *I-incorrect* (Table 1 row 2). With these statistics, we estimate 67% user input to have received reasonable responses and another 19% answered with uncertainty. This is comparable to, if not better than, performances reported in several studies of CAs [**?**, **?**] and commercial chatbots [**?**].

Developing the NLCs required an intent classifier schema and training data for each intent class. To construct the schema, we adopted an iterative data-driven approach by: a) content analysis on data from pilot studies; and b) extracting frequent topics from anonymized inquiry emails sent to an HR service center. With expert input from HR specialists, we grouped related NLC2 together to identify general intent categories as NLC1. With the schema, we obtained training data by extracting questions corresponding to each intent from data of pilot studies and HR inquiry emails through a text analytic process. Additional training examples were manually put in wherever necessary. We trained the classifier and developed the CA using IBM Watson Dialog. The curated answers for each NLC2 intent were either extracted by a text analytic process with the HR inquiry emails or manually created.

METHODOLOGY

Deployment and Participants

We recruited 337 participants through HR contacts in three groups with different starting dates, each using Chip for 5-6 weeks. Participants were college new hires with diverse backgrounds, joining different departments including engineers, consultants, designers, etc. They were located in multiple areas in the United States, including California, Massachusetts, Texas, etc. Upon joining, they attended an orientation session, where Chip was introduced by members of the research team. The introduction included a demo of Chip conversations, an overview of its functions for company related information assistance, suggestion to use *#fail* to provide feedback, and consent for the user study. The participation in the study was voluntary and no financial incentive was provided.

[2]In incorrect intent recognition, Chip either retrieves answers from the search engine, if available, or outputs the answer linked to NLC2

Survey

Within one week after the deployment period, we sent out a survey to participants' work emails. The survey response rate was 34.1% (N=115). We aimed to capture participants' self-reported satisfaction with Chip's functional performance and playful interactions to be used as ground-truth for studying predictive signals in conversational interactions. For functional satisfaction, given the targeted function of Chip as a QA system, we measured it by three items and used the average ratings to represent individual users' satisfaction with the instrumental usage of Chip (Cronbach $\alpha = 0.86$):

- *Understanding*: Chip was able to understand my questions.
- *Relevance*: Chip was able to find relevant information.
- *Quality*: The answers Chip provided were high quality.

In the background section, we discussed that playfulness implies two intertwining aspects of interactions: 1) engaging in human-like social conversations; 2) actively conversing about topics outside the task domain [**?**, **?**]. We adapted the agent sociality scale from our previous work [**?**] and asked participants to rate the following two items. We used the average ratings to represent individual users' satisfaction from playful interactions with Chip (Cronbach $\alpha = 0.73$):

- *Sociability*: I enjoyed chatting casually with Chip.
- *Off-topic desirability*: I enjoyed talking to Chip about topics unrelated to IBM knowledge and process.

All items used 7-point Likert-scales. We note that evaluation measure is still a challenge in studying CAs. Conventional usability scales appear to be less applicable to agents (e.g., efficiency). Most studies in this field (e.g. [**?**, **?**, **?**, **?**]), as ours, used homegrown questionnaires to capture user opinions.

Preparing for analysis: conversational acts labels

For RQ1, we aim to present a characterization of users' conversational input. Instead of looking at low-level utterances such as a user saying "*hi*" or "*hello*", we focus on higher-level conversational acts [**?**, **?**], which describe the performative functions of utterances in communication processes. That is, we are interested in how often users started with *greetings*, which would encompass many forms of utterances (*hi*, *hello*, *good morning*, etc.). We therefore resorted to a group of high-level NLCs (NLC1), as discussed in the methodology section, concerning with *conversational interactions*, defined as *communicative or social utterances outside work related QAs*. Specifically, these NLC1 classes include opening, closing, compliment, acknowledging, complaints, feedback (#fail), agent status chitchat, agent trait chitchat, agent ability checking, off-topic request, about me, and emoticon. Table 3 gives examples of sub-categories (NLC2) for each of them. To distinguish them from task related QA intents, we will refer to these NLCs as *conversational acts* in the rest of the paper.

As shown in the third row of Table 1, more than 80% of user input classified to be conversational acts by the NLCs fell in the correct intent recognition category, so we may use these labels with confidence. We manually went through the 471 user input classified to be conversational acts but fell in either incorrect intent recognition or low confidence category. We

	N	*Start day*	*Days of use*	*Mean (msg)*
Group 1	89	1	42	15.6
Group 2	121	15	36	20.9
Group 3	127	29	36	16.5

Table 2. Usage information of the three cohorts of participants

identified 233 to be conversational and corrected the labels, wherever necessary, then included them in the analysis. For the rest of the paper, when discussing conversational acts, we will refer to the NLC1 labels of these included cases.

RESULTS

To begin with, we report descriptive statistics of usages of Chip ($N_{user} = 337$, $N_{messages} = 6,004$ [3]). Table 2 shows the statistics of the three groups of participants. Figure 2 (top) shows the number of total messages by day for each group. Usages were most active in the first two weeks after the introduction, but remained substantial (day 1 was a Wednesday, and usage was low during weekends). We note that HR information assistance should have naturally undergone declining needs as new hires adapted to the work environment. Given our focus on conversational interactions, we examined the temporal patterns of them. Percentage of conversational interactions is calculated by the number of messages classified to be conversational acts divided by the total number of messages (QA+conversational interaction). As shown in Figure 2 (bottom), the percentages were generally consistent over the whole period. This is an important observation, suggesting that conversational interactions were not phenomenon due to novelty, but a regular part of user interactions with Chip.

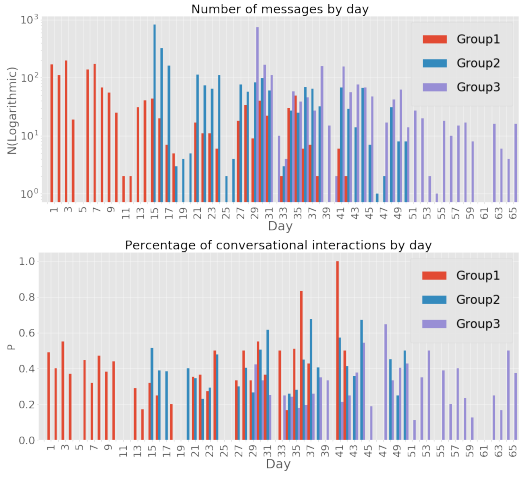

Figure 2. Number of user messages by day (top) and percentage of conversational interactions (bottom)

In the following, we first report on the patterns of users' conversational interactions with Chip (RQ1), then explore what signals exist in the conversational interactions for inferring user satisfaction with Chip's functional performance and playful interactions (RQ2) through the lens of statistical modeling.

Patterns of conversational interactions (RQ1)

Table 3 presents occurrence statistics of the 12 main categories of conversational interactions, specifically, the percentage of messages in each conversational act category (*P(msg)*)

[3]Except when Chip sent reminders and performing search, user input and Chip response happened strictly in pairs

Conversational NLC1	**Top NLC2**	**P(msg)**	**P(user)**
OPENING	Hello, Good morning, Are you there	7.7	57.3
CLOSING	Bye, I have to go, Nothing else now	0.9	11.6
COMPLIMENT	That is great, You are good/ smart/ cool/ helpful	1.1	11.9
ACKNOWLEDGE	Thanks, User acknowledge (*ok*, *got it*), User forgiveness (*no worries*)	6.0	46.6
COMPLAINTS	Wrong answer, Shut up, You are stupid	2.1	21.1
FEEDBACK	#fail feedback	7.7	42.4
AGENT STATUS CHITCHAT	How are you, What are you doing	1.7	20.2
AGENT TRAIT CHITCHAT	What do you like, What is your favorite, Agent identity, Agent age	2.7	22.3
AGENT ABILITY CHECK	What can you do, Can you do [function], How do you learn	1.8	22.0
OFF TOPIC REQUEST	Deliver food, About love, Meaning of life, Tell me a joke	3.2	27.0
ABOUT ME	Knowledge about me (*who am I? what do you know about me?*)	0.7	7.7
EMOTICON	-	0.7	7.1
All conversations		36.4	84.9

Table 3. Conversational acts NLCs and statistics: the percentage of each category over all user messages (*P(msg)*) and percentage of users had the category of conversational acts (*P(user)*)

(divided by the total number of messages), and the percentages of users who ever sent out such conversational messages (*P(user)*). Overall, despite its targeted usage as a QA system, a vast majority of users (84.9%) engaged in some forms of conversations. In total, conversational interactions accounted for 36.4% of user utterances. This highlights the importance of anticipating and designing to support users' interest areas in conversing with CAs. Among all conversational acts, the most frequent were chat opening (e.g., "*hi Chip*"), giving feedback using *#fail*, acknowledging message ("*ok*", "*thanks*"), followed by off-topic request (defined as request unrelated to work, such as "*tell me a joke*"), agent trait chitchat (e.g.,"*what do you like?*"), agent ability checking (e.g., "*what can you do?*", "*can you do [function] ?*"), complaints (e.g.,"*shut up*") and agent status chitchat (e.g., "*how are you?*","*what are you up to?*"). Comparatively, closing (e.g.,"*bye*)", about me (e.g., "*who am I?*"), and emoticon were less common categories.

Main areas of conversational interactions

While the application context might render idiosyncrasies to the interactions, we compared the observations to conversational interactions reported in previous studies of CAs in an attempt to identify common patterns. As mentioned, there were only a small number of studies providing schemes of conversational acts with CAs [**?**, **?**, **?**]. Among them, [**?**] reported the richest set of conversational acts, where museum visitors interacted with Max, an embodied CA providing information about the museum. Despite differences in the embodiment, Max shares key similarities with Chip, as both are to support domain specific QA, allow free-form user input, and are designed with capabilities to handle common chit-chat. Therefore, we chose Max to conduct a close comparison, based on the conversation schema given in [**?**]. It revealed many similarities in the types of conversational acts occurred. Based on our observations corroborated by those of Max, we summarize four main areas of interest in user initiated conversational interactions with QA agents below.

Feedback giving: In both cases, there was significant amount of feedback for agent's preceding responses. We observed similar occurrence percentages of compliment ("positive feedback" in [**?**]). In case of unsatisfied responses, we asked participants to use "#fail" to give negative feedback. 42.4% of users did it at least once. With this, we saw significant fewer blunt complaints with Chip compared to Max (a sub-category under "flaming" in [**?**]). Such frequent voluntary feedback has important implications for advocating conversational interface for information systems, given that it is a known challenge to elicit real-time feedback from users. In the next section, we study what kinds of conversations can be used as reliable signals for user feedback on the system performance.

Chit-chat about the agent: These take a significant portion of conversational interactions. Comparing Max and Chip, we observed close occurrence percentages of agent status chitchat ("anthropomorphic questions" in [**?**]), agent trait chitchat ("question about Max" in [**?**]) and off-topic request ("testing system" in [**?**]) . We will examine whether they reliably signal playful intentions in the next section.

Agent ability checking: This is a unique category we observed in interactions with Chip, such as asking "what can you do?" or "can you do [function]?". While we cannot conclude whether this was not identified to be a separate category in [**?**], we highlight that according to [**?**], when interacting with task-oriented agents, ability checking may carry distinct meaning from other anthropomorphic inquiries about the agents, as in directly serving the goal of understanding and reducing uncertainty about the system.

Communicative utterances: We observed frequent utterances that are habits in human communication process. One example is acknowledging (e.g.,"ok", "got it"), which happened much more frequently with Chip than with Max, potentially because they are habits of using the familiar chat interface. Opening also happened more frequently with Chip but the percentage of users who had it was similar to that of Max. This was potentially because users had repetitive, but shorter interaction sessions with Chip. We found a lower percentage of users explicitly closing the conversations with Chip. The reason could be that users tended to directly close the chat window. In the next section, we will examine whether these behaviors are habitual utterances or carry conscious playful intention by anthropomorphizing the agent.

Conversations as signals of user satisfaction (RQ2)

After providing a characterization of conversational interactions, we study what signals exist in them for inferring user satisfaction with Chip's functional performance and playful interactions. Figure 3 presents the distributions of user ratings on functionality and playfulness. It shows that there were fairly divided opinions on both Chip's functionality and playfulness, and the two aspects had only moderate correlation ($r(115) = 0.41$), highlighting the needs for adapting both system functions and interaction styles for different users. Satisfaction with playfulness is almost uniformly distributed. Given that Chip was designed with the capabilities to handle substantial amount of conversational interactions, we interpret it as individual difference in the tendency to seek satisfaction from playful interactions. This is consistent with the notion

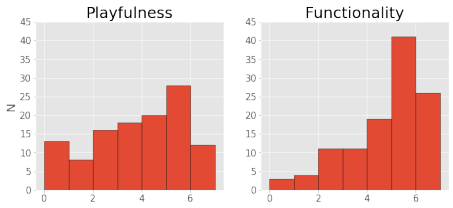

Figure 3. Distribution of survey ratings

of playfulness as an individual characteristic [**?**], and echoes conclusions from previous work [**?**, **?**] that user orientation for playful interactions may particularly merit system adaption. We used the self-reported ratings on functionality and playfulness of Chip as ground truth (*dependent variables*) to study predictive signals in user interactions. In the following, we present the statistical models and discuss the implications of the results for obtaining signals for user satisfaction.

Statistical model and results

A key insight from Cassell et al.'s work is that signals of interlocutors' positivity status exist in conversational behaviors of varied levels of granularity, from lexical choices, conversational acts to communication strategies [**?**, **?**, **?**]. We are interested in which conversational acts (*C-act*) can be used as reliable signals for user satisfaction with Chip. We also explore what lexical features (*Lex*) representing different conversational behaviors can provide additional signals. We note that our goal is not to build high-performance predictive models, but rather, through modeling of empirical data, we aim to identify relatively strong associations between certain conversational behaviors and users satisfaction.

We considered two types of features as independent variables (predictive features): conversational acts and lexical features representing conversational behaviors. For conversational acts (*C-act*), we included the occurrences of the 12 categories listed in Table 3 in each participant's interactions, normalized by his or her total number of messages. For lexical features (*Lex*), we performed a feature selection process on top of the standard bag-of-words method. The rationale is that we were interested in *conversational behaviors* instead of the content. So we were not interested in whether someone asked about healthcare or IT, but the way someone asking it (e.g., more formally "*how can I find...*" versus "*tell me...*") might be of our interest in differentiating conversational behaviors. Given our limited data size with survey responses ($N = 115$), the selective process also aimed to guard against overly sparse models that would endanger the validity of our conclusions.

Specifically, we started by converting all texts to lowercase and removed punctuations, then extracted trigram "bags of words" from messages of each participant to represent his or her lexical features—i.e., the occurrences of single words (unigrams), two-word (bigrams) and three-word (trigrams) phrases. To avoid sparsity, we kept only ones that 5% or more users used and had more than 0.1% occurrence in the whole message corpus ($N = 245$). To exclude domain specific content, we reviewed the extracted words and identified 46 entities to be specific to the HR domain (e.g. insurance, email, expense report, names of internal IT systems, etc.). We removed lexical features containing them ($N = 163$). Following conventions, we removed unigrams that were stop words (e.g.,

Functionality				Playfulness			
F	β (C-act)	β (C-act+Lex)	*F (Lex)*	*F*	β (C-act)	β (C-act+Lex)	*F (Lex)*
AGENT ABILITY CHECK	−.23	−.20	*what does* (.14), *tell me*	**AGENT STATUS CHITCHAT**	.18	.05	*how do you* (.24),
#FAIL	−.16	−.18	*about* (.11), *should I*	**COMPLIMENT**	.14	.08	*information* (.23), *should*
CLOSING	−.05	−.06	(.09), *ok* (.08), *where*	**AGENT TRAIT CHITCHAT**	.12	.07	*I* (.20), *I have* (.18),
OFF TOPIC REQUEST	−.02	−.02	*is/are* (.06), *who is/are*	**ABOUT ME**	.12	.09	*search* (.15), *is/are your*
ctr.: success rate	.19	.21	*my* (.03), *how to* (.02),	ctr.: success rate	.07	.13	(.12),*how are you* (.10),
ctr.: N(msg)	.07		*hi* (.02), *information*	ctr.: N(msg)	.02		*thanks* (.05),*tell me*
ctr.: Duration)	.11	.10	(.02), *how do I* (.02),				(.01),*do you know* (.01),
Df	7	18	*search* (-.02), *can you*	Df	6	18	*what can I* (-.04), *where*
%Dev	23.0	31.5	*do* (-.03)	%Dev	12.1	30.2	*do I* (-.04), *how do I* (-.06)

Table 4. Coefficients in Lasso regression models. Two models (*C-act* only and *C-act+Lex*) are presented for predicting each user satisfaction aspect—functionality and playfulness. The magnitude of a coefficient indicates the predictive power of the feature. Bold ones are predictive conversational acts. The last columns show the predictive lexical features, with coefficients in parenthesis. ctr. means a control variable.

to, at, the) and common verbs (e.g., go, find, look) since they were less interesting to represent language behaviors, as well as emoticons and the word "*fail*" since they were counted in conversational acts. We ended up with 76 lexical features.

We included three features as control variables for each individual: *total number of messages*, *duration of use*—calculated by the days between the first and last message, and *success rate*—as a proxy, ratio of input falling in the "correct intent" category (Table 1) to control for the system performance. Controlling for system performance is a critical consideration because we are interested in the variations in the *subjective* opinions. Implicit feedback is useful for identifying user groups that are less satisfied with *given system performance* to accommodate their different information needs or expectation.

We employed a penalized linear regression known as Lasso regression to model what conversational features predict user satisfaction with functionality and playfulness of Chip, respectively. Regular regression model performs best when features are independent of each other, so collinearity presents an issue. However, phrase collinearity is a known property of natural language. For example, the words "Chip" and "you" tended to co-occur in chitchat with Chip. With regular regression one would have to exclude either. Penalized regression model is known to guard against feature collinearity and sparsity, which works by shrinking coefficients of unimportant features to zero, leaving only reliably predictive features in correlated clusters. Therefore, Lasso regression has been frequently used to model language behaviors (e.g., [**?**, **?**, **?**]).

We ran Lasso regressions with the *glmnet* package of R. In Lasso regression, a parameter λ can be tuned to decide how much the model fits the data. A common problem with a large number of features is over-fitting—the statistical model fits the data too closely but sacrifices the general validity. We used the R package *cv.glmnet* to run cross-validation, a common technique to guard against over-fitting, by building a model on random sub-sets of the data and selecting λ that best predicts the rest. We built two regression models, one with conversational acts only (*C-act*) and one with additional lexical features (*C-act+Lex*). All dependent and independent variables were normalized. Unlike regular regression, Lasso does not provide significance tests (p-values) given that only predictive features are kept in the model. We examined the β coefficients of variables in the selected model (with the optimal λ) to determine the relative predictive powers. In Table 4, we present β coefficients of predictive conversational acts features in both models (*C-act*, *C-act+Lex*), and the additional predictive lexical features in the last columns. Using these model results as a lens, below we discuss areas to obtain signals for user satisfactions with CA functionality and playfulness.

On instrumental usage satisfaction

Opportunities and caveats in conversational feedback: Giving *negative feedback* using "#fail" signals negative opinions on functionality after controlling for the system performance, meaning participants who had lower *subjective* satisfaction were more likely to give such feedback. These are the users that a system should attend to for potentially different information needs. Considering that more than 42% users provided such signals, it highlights the benefit of conversational interfaces for obtaining real-time feedback. In interesting contrast, *positive compliment* such as "you are smart" predicts playfulness instead of positivity on task performance, so they may not be taken as reliable signals for user feedback on the instrumental usage. It is worth pointing out that complaints did not appear to be predictive. In our case, they happened sparsely as we asked participants to use "#fail" to express dissatisfaction. It suggests that, to obtain reliable feedback for algorithmic improvement, it may be beneficial to ask users to provide it in a standard form. For example, Google Allo provides thumbs-up and -down buttons for giving feedback.

Implicit complaints: The occurrences of *agent ability check* and *closing* negatively predict the satisfaction on functionality. One explanation is that ability-check might be resulted from frustration, as we observed user asking "*what can you do?*" or "*can you do ...?*" after encountering errors. [**?**] discussed a key challenge in using CAs being "gulf of execution"—unclear affordance. Agent ability-check can be considered signals of user struggling with the gulf of execution. Similarly, we observed users closing the conversation ("*bye*") after errors, signaling frustration and refusal to continue using the system. All predictive conversational acts for functionality have negative coefficients, suggesting that conversational interactions are in general a source for identifying user frustration .

Formal QA as signals of functional satisfaction: A negative signaling effect of off-topic requests such as "tell me a joke" was found for user satisfaction on functionality. Meanwhile, the positive lexical features for functionality indicate formal patterns of questioning (*what/ where/ who/ how*). This is to be expected, as those satisfied with Chip's performance were likely to continuously use it for serious information needs. It suggests a simple way to infer functional satisfaction could be to monitor the frequency of formal QAs.

On being happily playful

Playful chitchat, not habitual utterances: Three categories of conversational acts are the strongest signals of playfulness—chitchat *asking about agent's traits*, *asking about agent's status*, and *talking about oneself*. They confirm that chit-chat carries explicit playful intentions. Although a previous study showed a correlation between greeting and a tendency to anthropomorphize an embodied agent [**?**], we found no signaling effect of chat opening, but the lexical feature "hi" predicts functional satisfaction, potentially indicating a higher tendency to open the chat for instrumental usage. A similar trend was found for the lexical feature "ok", a common acknowledging phrase after receiving messages. This confirms that, in the context of text-based QA agents, chat opening and acknowledging may be more of habitual utterances with the chat interface instead of consciously anthropomorphizing the agent.

Agent orientated conversations as seeking playfulness: An evident pattern in lexical features signaling playfulness is frequent occurrences of second-person pronouns (e.g., "*how do you*", "*how are you*") in positive features and first-person pronouns in negative features. This agent oriented interest in conversations is consistent with the tendency to anthropomorphize the agent and engage in chit-chat. This suggests a simple way to identify playful users could be monitoring the usage of second-person pronouns, which was studied in HRI work as signals for social interaction inclination with robots [**?**].

Casual testing as seeking playfulness: In contrast to functionality, the lexical features predicting satisfaction from playfulness suggest less formality such as "*how do I*" and "*what can I*", but more casual asking such as "*do you know*" or "*tell me*". We also found the words *information* and *search* to be strong signals for playfulness. A close examination of the actual conversations revealed a pattern of repeatedly asking Chip to retrieve different kinds of information (e.g.,"*search information about my manager*"). Consistent with the definition of computer playfulness [**?**], these behaviors suggest "testing intelligence" to be a manifestation of playfulness.

To summarize, we found that there are reliable signals in conversational interactions for inferring user satisfaction. By contrasting signals for instrumental usage versus playful interactions, we further shed light on the varying functions underlying different conversational behaviors. Revisiting the four interest areas of conversations identified in the last section, we may conclude that: 1) Not all feedback is regarding the instrumental usage of CAs. Some may be exhibited with playful intentions. But user behaviors can be regulated to obtain reliable feedback; 2) Agent oriented chit-chat is indeed an indication of users seeking satisfaction from playful interactions; 3) Agent ability checking can be considered signals of users struggling with the system's functional affordance; 4) Communicative utterances such as opening and acknowledging are more of habitual behaviors in using the chat interface.

DISCUSSIONS

Conversational interactions

Despite the system working as a QA agent, conversational interactions outside querying information were common. These observations are indicative of the central interest areas of user initiated conversations with QA agents, as corroborated by previous studies [**?**, **?**, **?**]. Several take-aways may inform future development of CAs. First, users actively engaged in explicit feedback-giving, and more implicit signals for user frustrations were also observed. Utilizing these behavioral signals may open up possibilities for building adaptive systems. Second, a challenging task in the development of CAs is to anticipate chit-chat in a free-form input. Our results suggest that a large proportion of it may be centered around the traits and status of agents. Designing an agent with a comprehensive and consistent persona and elaborating on descriptions of such a persona (e.g., what does the agent like) may help prepare for addressing this type of chit-chat. Lastly, habitual chatting behaviors such as opening and acknowledging were common, as invited by the familiar text-based chat interface. Content developers should anticipate habitual behaviors in similar human-human communication channels or contexts.

By contrasting conversational behaviors signaling instrumental usage and playful interactions, we illustrated that while some were consciously anthropomorphizing the agent, some were better seen as system operations in a conversational form. [**?**] made a distinction between relational and grounding conversations with CAs. The latter, including acknowledging and repair, are concerned more with achieving task goals. Sometimes the distinction may be less clear. For example, we found that the frequent question "what can you do" although on the surface was asked in an anthropomorphic way, served more for the instrumental usage of the system, in an analogy to visiting the "about" or "help" sections on a graphic interface. We also observed users asking Chip whether it could perform some advanced assistance. According to many, a limitation of CA interfaces is its ambiguity in affordance [**?**], creating "gulf of execution and evaluation" [**?**]. It is critical for designers to anticipate this type of user inquiries and carefully consider different entry points where users may seek information about the system affordance and usage.

With these observations, we revisit the idea of developing task-independent modules for conversational behaviors to be reused for developing CAs. For example, [**?**] proposed an architecture that provides a task-independent framework on agents' conversational strategies, including clarification, confirming and controlling turn-taking. However, its focus is on agent-initiative systems. Our results point to the possibility of a reusable conversational module for QA agents. We envision it to be equipped with generic, reusable responses to common communicative utterances, as well as example data and guidelines for developers to anticipate chit-chat and system inquiries from users, where the responses can be customized.

Inferring instrumental satisfaction from conversations

Although our current system is a static one, building adaptive agents that can accommodate different user needs is our ultimate goal and a longstanding interest of the research community. We highlight the necessity for identifying reliable signals to infer user status through analysis of human-agent interaction data. Although understanding how people manifest internal status in communicative behaviors with human partners is important for building realistically social agents [**?**, **?**], the deviations in conversations with Chip were evident.

In the less anthropomorphic context of QA agents, the human concept of social positivity may become obsolete. Our study shows that there are also rich signals in conversational behaviors to infer user satisfaction with the functional performance of CAs, especially for signaling frustration. Compared to query-based information systems, users are more likely to reveal their emotional status in conversational interactions to provide real-time feedback. This further highlights the importance of designing to support conversational interactions. Without designing proper responses (so users may stop after a few attempts), or by constraining user initiatives (e.g., button-based input), one may miss capturing these signals.

These signals can complement the limited set of "implicit feedback" used in IR systems to continuously monitor user satisfaction while optimizing algorithmic performances and system functions [**?**, **?**]. Providing agent-initiated assistance such as giving example questions or suggesting advanced functions to help those exhibiting frustration is another area to explore. Some adaption may also target users with high satisfaction. For example, [**?**] proposed a framework of bootstrapping algorithms by targeting "low-risk" users, who are currently at a high satisfaction level and are likely more forgiving for system failures or the cost of switching to new designs.

Agent playfulness

Previous work on the early adoption of personal computers considered playfulness as a desirable characteristic in computer interactions because it promotes adoption, satisfaction and learning outcome [**?**]. In particular, Webster et al. advocated playfulness in the workplace to make "*employees experience more positive affect at work.*" As an enterprise tool, Chip has the potential to fulfill such a function. Specifically for CAs, playfulness was considered as a"point of entry" [**?**]. This puts our study into perspective as we observed substantial playfulness in the first 6 weeks of deployment. Although [**?**] suggested that playful behaviors may decline in the long run and future research should examine such a possibility, we emphasize that supporting playful interactions can enhance adoption [**?**], especially for individuals who value this unique offering of CAs. One may also explore designs that can promote and sustain playfulness in the long run. For designing computer playfulness, Webster offered several guidelines, including arousing curiosity, reserving uncertainty, encouraging creativity, and exhibiting simplicity [**?**]. Translating to agent designs, potential features to consider are proactive social interactions, manifesting personality, continuously revealing new features and responses, avoiding complexity, providing transparency and user control for system status. [**?**] also emphasized system robustness in the expectation of playful users. This is underscored in our observation that playfulness is also manifested in actively testing the agent.

Identifying individual or situational differences in playfulness has been an interest for system adaption, which motivated the development of survey scales on computer playfulness [**?**]. Our results point to some easy-to-obtain signals for playfulness in interacting with CAs, from chit-chat, casual testing behaviors and lexical choices, to as simple as monitoring the use of pronouns. By detecting these signals, a CA may adapt its interaction style. Previous studies reported qualitative findings that playful users of CAs look for "humanized and humorous responses", "finding Easter eggs" [**?**], and "subjective opinions and personality" [**?**]. A caveat discussed in [**?**] is that agent responses for playful interactions may also serve as affordance cues, where over boasting of "social smarts" can belie the true system capabilities and raise incorrect user expectations. So it is an intricate task to design agent responses for playful interactions. One should leverage such user engagement to better support the usage of the agent. For example, it may be an opportunity to communicate about the system's functional scope when users initiate chitchat about agent's traits and status. Adaption for the opposite direction of playfulness may be equally important. Our previous study showed that some users have little interest in playful interactions with CAs and tend to exhibit utility-oriented behaviors and preferences, such as typing only keywords, and desiring responses resembling search results instead of lengthy conversations [**?**].

LIMITATIONS

We acknowledge several limitations. First, our results are based on survey data. Although 34.1% should be seen as a considerable response rate given the professional context, and there is a wide distribution of interaction frequencies among those responded, we cannot rule out self-selection bias. However, the focus of this paper is not to demonstrate the positivity of user opinions on Chip but build predictive models, so it should be less subject to the problem of self-selection bias. Second, as with any automatic methods to characterize large quantity of texts, the conversational labels we obtained were not without noise. The schema we used might not capture rarer cases of conversational interactions. However, our goal is not to establish a formal taxonomy but to contribute empirical insights by quantifying available types of conversational acts. Lastly, we acknowledge that some observations may be specific to the workplace context and user sample of the study, as young professionals may be more inclined for playful interactions. We do not claim the generalization of specific statistics but focus on the patterns they represent.

CONCLUSION

By studying log data from a field deployment of a question-and-answer conversational agent, we characterize the rich forms of conversational interactions users had with the agent. The main areas of conversations include feedback-giving, playful chit-chat, system inquiry, and habitual communicative utterances. Through the lens of statistical modeling, we highlight the rich signals in conversational interactions for inferring user satisfaction, which can be utilized to develop agents that can adapt algorithmic performances and interaction styles. The results also provide nuanced understanding on the underlying functions of conversational behaviors with QA agents and their deviations from human conversations. Our findings may inform designs of CAs and contribute to the emerging fields of conversational UX, conversational IR and adaptive agents.

ACKNOWLEDGEMENT

We appreciate the valuable input from anonymous reviewers. We would like to thank Thomas Erickson for providing feedback on the manuscript and all study participants.

REFERENCES

1. John Langshaw Austin. 1975. *How to do things with words*. Oxford university press.
2. Niels Bernsen and Laila Dybkjær. 2004. Domain-oriented conversation with HC Andersen. *Affective Dialogue Systems* (2004), 142–153.
3. Yulong Bian, Chenglei Yang, Dongdong Guan, Sa Xiao, Fengqiang Gao, Chia Shen, and Xiangxu Meng. 2016. Effects of Pedagogical Agent's Personality and Emotional Feedback Strategy on Chinese Students' Learning Experiences and Performance: A Study Based on Virtual Tai Chi Training Studio. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 433–444.
4. Timothy Bickmore, Laura Pfeifer, and Daniel Schulman. 2011. Relational agents improve engagement and learning in science museum visitors. In *International Workshop on Intelligent Virtual Agents*. Springer, 55–67.
5. Timothy W Bickmore, Laura M Pfeifer, and Brian W Jack. 2009. Taking the time to care: empowering low health literacy hospital patients with virtual nurse agents. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1265–1274.
6. Dan Bohus and Alexander I Rudnicky. 2003. RavenClaw: Dialog management using hierarchical task decomposition and an expectation agenda. (2003).
7. Susan E Brennan. 1990. Conversation as direct manipulation: An iconoclastic view. (1990).
8. Hendrik Buschmeier and Stefan Kopp. 2011. Towards conversational agents that attend to and adapt to communicative user feedback. In *Intelligent Virtual Agents*. Springer, 169–182.
9. John M Carroll and John C Thomas. 1988. Fun. *ACM SIGCHI Bulletin* 19, 3 (1988), 21–24.
10. Justine Cassell. 2001. Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine* 22, 4 (2001), 67.
11. Justine Cassell and Timothy Bickmore. 2003. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User modeling and user-adapted interaction* 13, 1 (2003), 89–132.
12. Praveen Chandar, Yasaman Khazaeni, Matthew Davis, Michael Muller, Marco Crasso, Q Vera Liao, N Sadat Shami, and Werner Geyer. 2017. Leveraging Conversational Systems to Assists New Hires During Onboarding. In *IFIP Conference on Human-Computer Interaction*. Springer, 381–391.
13. Mark Coeckelbergh. 2011. You, robot: on the linguistic construction of artificial others. *AI & society* 26, 1 (2011), 61–69.
14. Mark G Core and James Allen. 1997. Coding dialogs with the DAMSL annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, Vol. 56. Boston, MA.
15. Doris M Dehn and Susanne Van Mulken. 2000. The impact of animated interface agents: a review of empirical research. *International journal of human-computer studies* 52, 1 (2000), 1–22.
16. Nicholas Epley, Adam Waytz, and John T Cacioppo. 2007. On seeing human: a three-factor theory of anthropomorphism. *Psychological review* 114, 4 (2007), 864.
17. Dan Fletcher. 2010. The 50 Worst Inventions: Microsoft Bob. TIME. (27 May 2010).
18. Eric Gilbert. 2012. Phrases that signal workplace hierarchy. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 1037–1046.
19. Joakim Gustafson and Linda Bell. 2000. Speech technology on trial: Experiences from the August system. *Natural Language Engineering* 6, 3-4 (2000), 273–286.
20. John Heritage and John Maxwell Atkinson. 1984. *Structures of social action: Studies in conversation analysis*. Cambridge University Press.
21. Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic online evaluation of intelligent assistants. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 506–516.
22. Diane Kelly and Jaime Teevan. 2003. Implicit feedback for inferring user preference: a bibliography. In *ACM SIGIR Forum*, Vol. 37. ACM, 18–28.
23. Alfred Kobsa and Wolfgang Wahlster. 1989. *User models in dialog systems*. Springer.
24. Stefan Kopp, Lars Gesellensetter, Nicole C Krämer, and Ipke Wachsmuth. 2005. A conversational agent as museum guide–design and evaluation of a real-world application. In *International Workshop on Intelligent Virtual Agents*. Springer, 329–343.
25. Min Kyung Lee, Sara Kiesler, and Jodi Forlizzi. 2010. Receptionist or information kiosk: How do people talk with a robot?. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 31–40.
26. Namseok Lee, Hochul Shin, and S Shyam Sundar. 2011. Utilitarian vs. hedonic robots: role of parasocial tendency and anthropomorphism in shaping user attitudes. In *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*. IEEE, 183–184.
27. Q Vera Liao, Matthew Davis, Werner Geyer, Michael Muller, and N Sadat Shami. 2016. What Can You Do?: Studying Social-Agent Orientation and Agent Proactive Interactions with an Agent for Employees. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. ACM, 264–275.

28. Ewa Luger and Abigail Sellen. 2016. Like Having a Really Bad PA: The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5286–5297.
29. Chris Matyszczyk. 2012. Apple's Siri wrong 38 percent of the time in test. CNET. (30 June 2012).
30. Tanushree Mitra and Eric Gilbert. 2014. The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 49–61.
31. Robert J. Moore, Rafah A. Hosn, and Ashima Arora. 2016. The Machinery of Natural Conversation and the Design of Conversational Machines. In *American Sociological Association annual meeting.*
32. Amy Ogan, Samantha Finkelstein, Elijah Mayfield, Claudia D'Adamo, Noboru Matsuda, and Justine Cassell. 2012a. Oh dear stacy!: social interaction, elaboration, and learning with teachable agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 39–48.
33. Amy Ogan, Samantha L Finkelstein, Erin Walker, Ryan Carlson, and Justine Cassell. 2012b. Rudeness and Rapport: Insults and Learning Gains in Peer Tutoring.. In *ITS*. Springer, 11–21.
34. Christopher Peters, Stylianos Asteriadis, and Kostas Karpouzis. 2010. Investigating shared attention with a virtual agent using a gaze-based interface. *Journal on Multimodal User Interfaces* 3, 1 (2010), 119–130.
35. Martin Porcheron, Joel E Fischer, and Sarah Sharples. 2017. Do Animals Have A ccents?: Talking with Agents in Multi-Party Conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 207–219.
36. Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. ACM, 117–126.
37. Susan Robinson, Antonio Roque, and David R Traum. 2010. Dialogues in Context: An Objective User-Oriented Evaluation Approach for Virtual Human Dialogue.. In *LREC*.
38. Susan Robinson, David R Traum, Midhun Ittycheriah, and Joe Henderer. 2008. What would you Ask a conversational Agent? Observations of Human-Agent Dialogues in a Museum Setting.. In *LREC*.
39. Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323.
40. Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on*. IEEE, 305–311.
41. John R Searle. 1976. A classification of illocutionary acts. *Language in society* 5, 01 (1976), 1–23.
42. John R Searle, Ferenc Kiefer, and Manfred Bierwisch. 1980. *Speech act theory and pragmatics*. Vol. 10. Springer.
43. Alexander Serenko. 2008. A model of user adoption of interface agents for email notification. *Interacting with Computers* 20, 4-5 (2008), 461–472.
44. Nicole Shechtman and Leonard M Horowitz. 2003. Media inequality in conversation: how people behave differently when interacting with computers and people. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 281–288.
45. Ben Shneiderman and Pattie Maes. 1997. Direct manipulation vs. interface agents. *interactions* 4, 6 (1997), 42–61.
46. Yang Song and Li-wei He. 2010. Optimal rare query suggestion with implicit user feedback. In *Proceedings of the 19th international conference on World wide web*. ACM, 901–910.
47. Luke Swartz. 2003. *Why people hate the paperclip: Labels, appearance, behavior, and social responses to user interface agents*. Ph.D. Dissertation. Stanford University Palo Alto, CA.
48. Daniel Szafir and Bilge Mutlu. 2012. Pay attention!: designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 11–20.
49. Jaime Teevan, Susan T Dumais, and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 449–456.
50. Robert Trappl. 2013. *Your Virtual Butler*. Springer.
51. Marilyn A Walker, Diane J Litman, Candace A Kamm, and Alicia Abella. 1997. PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 271–280.
52. William Yang Wang, Samantha Finkelstein, Amy Ogan, Alan W Black, and Justine Cassell. 2012. Love ya, jerkface: using sparse log-linear models to build positive (and impolite) relationships with teens. In *Proceedings of the 13th annual meeting of the special interest group on discourse and dialogue*. Association for Computational Linguistics, 20–29.

53. Jane Webster. 1988. Making computer tasks at work more playful: Implications for systems analysts and designers. In *Proceedings of the ACM SIGCPR conference on Management of information systems personnel*. ACM, 78–87.

54. Jane Webster and Joseph J Martocchio. 1992. Microcomputer playfulness: Development of a measure with workplace implications. *MIS quarterly* (1992), 201–226.

55. Joseph Weizenbaum. 1966. ELIZAâĂŤa computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.

56. Jun Xiao, John Stasko, and Richard Catrambone. 2004. An empirical study of the effect of agent competence on user performance and perception. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 1*. IEEE Computer Society, 178–185.

57. Jun Xiao, John Stasko, and Richard Catrambone. 2007. The role of choice and customization on users' interaction with embodied conversational agents: effects on perception and performance. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 1293–1302.

58. ChengXiang Zhai and John Lafferty. 2006. A risk minimization framework for information retrieval. *Information Processing & Management* 42, 1 (2006), 31–55.

59. Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. 2016. Automatic Recognition of Conversational Strategies in the Service of a Socially-Aware Dialog System.