Exploring the Effects of Self-Mockery to Improve Task-Oriented Chatbot's Social Intelligence

Chengzhong Liu
chengzhong.liu@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Shixu Zhou
szhouav@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Yuanhao Zhang
yzhangiy@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Dingdong Liu
dliuak@connect.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

Zhenhui Peng
pengzhh29@mail.sysu.edu.cn
Sun Yat-sen University
Zhuhai, China

Xiaojuan Ma*
mxj@cse.ust.hk
The Hong Kong University of Science and Technology
Hong Kong, China

ABSTRACT

An effective task-oriented chatbot should be able to exert a certain level of Social Intelligence (SI), the ability to emulate human social behaviors to reduce user frustration and dissatisfaction. However, few studies explored using humor, a common rhetorical device in human-human interactions, to improve chatbots' overall SI. To fill this gap, we proposed to apply self-mockery humor to a customer service chatbot in different interaction stages with users. We proposed a pipeline to create situated self-mockery for the chatbot and conducted a within-subject experiment (N=28) to compare it with a chatbot without self-mockery utterance. Results showed that the self-mockery chatbot was perceived as significantly funnier, more satisfactory, and delivering higher performance in two out of the five measured characteristics of SI with comparable performance in the rest. We further discussed how participants' individual factors might affect the perceived helpfulness of self-mockery on SI and concluded with design considerations.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; **Empirical studies in HCI**.

KEYWORDS

Self-Mockery, Chatbots, Conversational Agents, Social Intelligence

1 INTRODUCTION

Task-oriented chatbots, have undergone fast development in recent years [6, 40] and are applied to numerous domains, including but not limited to education, customer service, and health [1]. They are generally developed to complete specific tasks with text-based input & output, in contrast to general-purpose chatbots designed for chit-chat and entertainment [13, 28]. However, as the chatbots are still inadequate at fully understanding the complex natural language by people, breakdowns [6] and avoidance [43] frequently happen during their interactions with human users. If not handled appropriately, such shortcomings could lead to user dissatisfaction and frustration, triggering abusive behaviors of users towards chatbots [4, 14, 15] or even abandonment of the conversation [38].

To mitigate the aforementioned challenges, Human-Computer Interaction (HCI) researchers have tried to improve chatbots' social intelligence, the ability to adapt human social behaviors [13]. Specifically, it refers to the ability of an intelligent agent to respond to social cues, accept differences, manage conflicts, be empathic and demonstrate caring behaviors during the conversation with users [9, 60], which can be ascribed to distinctive and measurable characteristics [13], including *damage control*, *thoroughness*, *manners*, *moral agency*, *emotional intelligence*, and *personalization*. To fulfill this goal, existing works in HCI have applied humor, a common way to reduce anxiety and adjust the tense social atmosphere in the Human-Human Interactions (HHI) [73, 79], to intelligent agents. This approach yielded fruitful benefits, e.g., making the agent more human-like to reduce the unpleasantness when the agent fails [46, 52] and gaining higher likability during the conversation [8, 67].

Nevertheless, these studies either focused on humor in non-task-oriented social agents [8, 67] or targeted usage of humor in a particular scenario of a task in HCI, such as humor in ice-breaking [67] or handling failure [46, 52]. Moreover, based on our survey, applying self-oriented humor, a.k.a., self-mockery, to task-oriented chatbots is understudied in the HCI domain, although in the HHI it has proven to be helpful for adjusting intense atmosphere in communication, alleviating mental pressure, and amending participants' relationships [45]. Hence, it is worth exploring the effects of applying self-mockery to task-oriented chatbots to exhibit overall social intelligence across multiple scenarios around a task and alleviate the challenges they face in more general settings.

To this end, we proposed designing a self-mockery generation pipeline for task-oriented chatbots and testing it with a simulated real-world use case. Different from previous studies that focused primarily on only one (e.g., *emotional intelligence* [43]) or a limited subset of these social intelligence characteristics [13, 21, 64, 71], e.g., using iterative development with limited topics [64], building vocabulary with specific community dialogues [62], and manifesting the particular characteristic(s) specific to certain scenarios (i.e., dealing

*Corresponding author

with abusive behaviors) [5], we attempted to develop a consistent mechanism for the chatbot to address the overall improvement of its social intelligence [13] which is potentially applicable to more general task scenarios. In this paper, we focused on exploring the following Research Questions (RQs):

- **RQ1** How to design self-mockery language for task-oriented chatbots, and how does it affect users' overall satisfaction?
- **RQ2** How do users perceive the overall effects of self-mockery and its helpfulness in different interaction stages towards task-oriented chatbots' social intelligence?
- **RQ3** How do users' individual factors influence their perception of the social intelligence of the self-mockery chatbot?

As users often consider chatbots as AI applications [29, 61], we followed the guideline of Human-AI interactions [2] to customize the self-mockery language to be applied in a variety of common scenarios during task-related conversations, namely, *greetings* [57], *user challenge* [43], and *handling failures* [6]. To test the efficacy of self-mockery in manifesting social intelligence in task-oriented chatbots, we implemented a customer service chatbot with the *RASA*[1] platform as a research probe. We conducted a within-subject experiment (N=28) to evaluate its perceived social intelligence compared to a baseline chatbot without self-mockery. Results showed that compared to the baseline chatbot, the self-mockery chatbot was perceived as significantly funnier, more satisfactory, and with higher performance in *damage control* and *emotional intelligence*; it has comparable performance with the baseline in the rest of the measured characteristics of the social intelligence. As for the participants' individual factors, only the participants with a higher social orientation towards chatbots [39, 40], i.e., preferring human-like chatbots more, would better appreciate the helpfulness of self-mockery on all characteristics of social intelligence in all scenarios. In contrast, other measured individual factors such as service orientation [37] and familiarity with the chatbot technology [6] did not appear to have a significant impact on this perception. Finally, we concluded with design considerations to design self-mockery for task-oriented chatbots. We summarized our contributions as:

- The design and evaluation of the pipeline to generate self-mockery language for task-oriented chatbots.
- Empirical insights from a within-subject study to understand the factors that influence the perceived social intelligence of the task-oriented chatbots with self-mockery design.

2 RELATED WORK

2.1 Social Intelligence of Chatbots

In recent years, text-based conversational agents, a.k.a., chatbots, have increasingly gained popularity and deployed to various fields [6, 38, 39]. With the rapid development of chatbot technologies, people increasingly regard chatbots as social actors, calling the need to develop chatbots' social intelligence to meet people's expectations better [59]. Chatbots with social intelligence can better deal with some of the existing problems and challenges, including improving the HCI experience, solving possible failure in the face of breakdowns, and avoiding falling into deadlock when the chatbot cannot understand the user's input and respond to it correctly [13].

[1]https://rasa.com/

According to the survey by Chaves et al., Social intelligence can be divided into six characteristics, *damage control*, *thoroughness*, *manners*, *moral agency*, *emotional intelligence*, and *personalization* [13]. *Damage control* refers to the ability of chatbots to deal with conflicts and failures [13]. *Thoroughness* is the ability of the chatbot to use language consistently and accurately [48]. *Manners* refer to whether the chatbot's dialogue habits in the conversation are polite [48]. *Moral agency* means whether the chatbot can have the correct view of right and wrong [7]. And *emotional intelligence* reflects the empathy and emotional expression ability of chatbot [60]. The final *personalization* represents whether the chatbot can self-adjust for different types of users and has the ability to serve different users and meet their preferences [22].

Many previous studies have focused on improving the social intelligence of a particular aspect of the chatbot. For example, Toxtli et al. designed a chatbot for task management and found that chatbot needs to know when to talk, which can make it behave in a more polite and human-like way as well as improve the chatbot's social intelligence of *manners* [68]. In the work of Wallis et al., they improved the chatbot's ability to handle users' inappropriate and aggressive input by allowing the chatbot to withdraw from the conversation and keep silent, which helps demonstrate the chatbot's social intelligence of *damage control* [71]. Kumar et al. successfully improved the chatbot's social intelligence of *emotional intelligence* by using solidarity, agreement, and tension release utterances in their proposed chatbot, and made it more likable and acceptable in group chats [35].

However, most of the aforementioned previous work only focused on limited characteristics of the chatbot's social intelligence, and few works studied how to improve multiple characteristics of the chatbot's social intelligence at the same time. Mariacher et al., whose work focused on both *emotional intelligence* and *thoroughness*, wrote text utterances based on social intelligence literature and used them as chatbot sentences to improve people's perception of the chatbot's social intelligence and the interpersonal human-chatbot interaction [44]. But in their work, they were trying to directly use the conversational language exhibiting social intelligence in the HHI as the template to write utterances for chatbots, which is not an easy way to implement and be applied to chatbots. Therefore, it is necessary to propose a new method that is more generalizable to improve chatbot's multiple characteristics of social intelligence all around.

2.2 Self-Mockery Language

Self-mockery is a form of humor that banters on oneself [45], which has been extensively studied and widely used in the HHI. Wen et al. argue that self-mockery could generally serve as a lubricant in social interactions and mediate the intense atmosphere [45]. A similar argument was also supported by Haugh's study, in which he proved that by self-deprecating (a.k.a., self-mockery), one could lower his/her place and position to reach a sense of "shared ordinariness" and promote communication and connection between strangers [25]. Also, Ungar et al. pointed out that self-mockery is a way of responding to the high expectations of others without losing dignity and politeness [69].

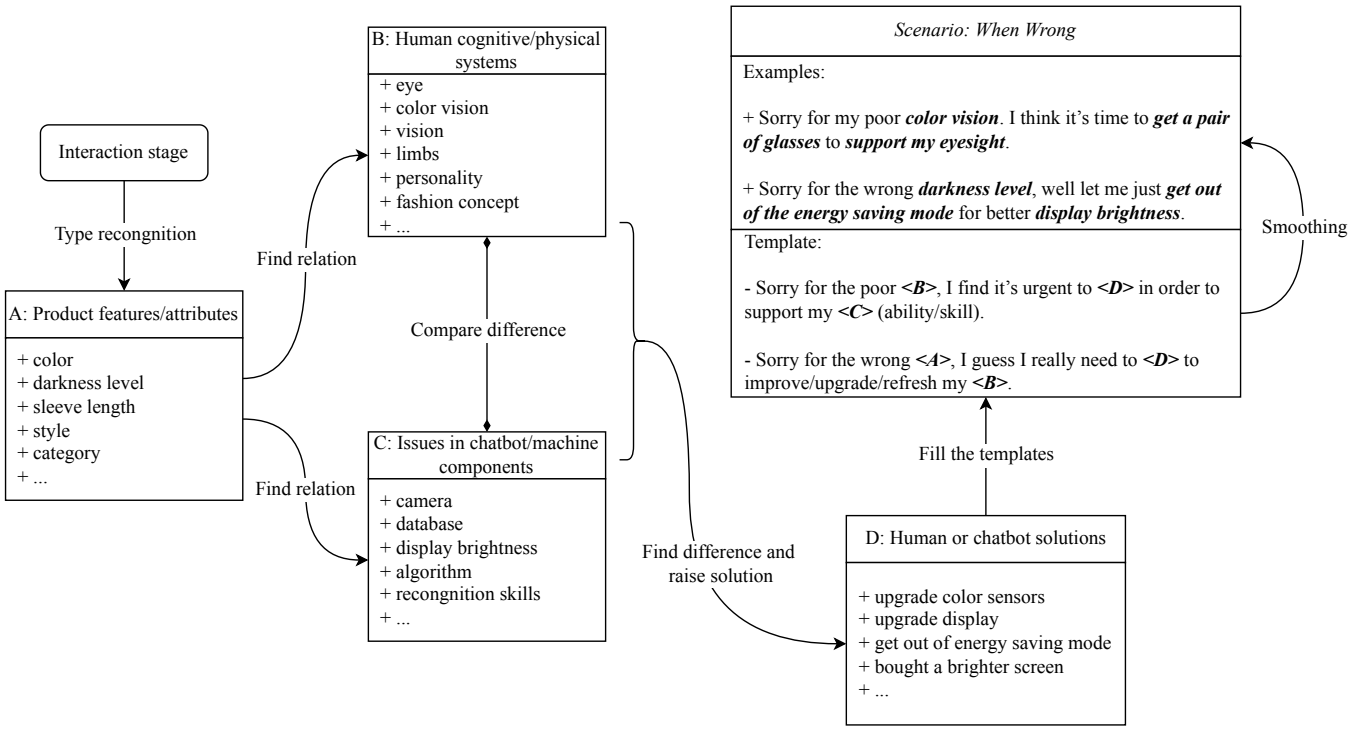

Figure 1: The self-mockery language generation pipeline for chatbots.

In the previous work of human-chatbot interaction, humor was widely used and improved the performance of chatbot perceived by users. For example, Tae et al.'s work used canned jokes as robot's ice-breaking sentences, and robot with humorous ice-breaking was perceived as more approachable for the users and gained higher likeability [67]. Moreover, Mulder et al. found that humor can make an agent (e.g., chatbots, robots) seem more human-like when it fails, which, in consequence, can ease the interaction [49].

However, in human-chatbot interaction, few works focused on investigating the benefits of applying self-oriented humor, a.k.a., self-mockery. Mirnig et al. used humor in robots and found that self-mockery forms of humor are rated significantly higher than aggressively schadenfreude ones when it comes to robot likeability, yet they only emulated the self-mockery of the robots via the sound of laugh [46]. Furthermore, most of the work related to humor in human-chatbot interaction only focuses on applying humor in a specific scenario, like when ice-breaking [67] and when chatbot fails [49]. Therefore, it is worth investigating the application of self-mockery humor into chatbots and a general design for self-mockery in various scenarios.

3 SELF-MOCKERY DESIGN

Following the guideline of Human-AI interactions summarized by [2], the interaction between user and chatbot can be concluded in four stages, namely, *initially*, *during interaction*, *when wrong*, and *over time*. Most of the commercially available task-oriented chatbots in business, such as *Bolt by Zoom*[2], are designed only to interact with the user to complete a task in a relatively short time [26, 51, 76]. We therefore did not include the interaction stage *over time* in our self-mockery design and correspondingly did not measure *personalization*, a characteristic of the chatbot's social intelligence generally manifested in long-term interactions [13, 39].

For the self-mockery language generation, we followed the previous survey [54] to explore both the Retrieval-Based method [77] and the Generation-Based method [63], two common methods that are pervasively used by the task-oriented chatbots [65]. After we attempted the two methods, we concluded that the Retrieval-Based method is not feasible to generate self-mockery for chatbots due to the sparsity of the appropriate self-oriented humor candidate sentences. Moreover, we could not find available Natural Language Processing (NLP) methods that can generate usable humorous utterances for chatbots on a large scale [3, 27, 36] to fulfill the demand of the Retrieval-Based method. Therefore, we adopted the Generation-Based method with designed templates and corresponding components.

3.1 Retrieval-Based Method

The retrieval-based method is to select an existing sentence for a chatbot in response to user input from a pre-compiled repository [54, 77]. Therefore to support the self-mockery utterances of the chatbot, we selected the *"Short-joke-dataset"*[3] as the candidate pool of the self-mockery. The dataset contains 2950 punchlines

[2]https://zoom.us/

[3]https://github.com/amoudgl/short-jokes-dataset/blob/master/data/onelinefun.csv

Table 1: Examples of self-mockery components selection for different scenarios: *Greetings*, *User Challenge*, and *Handling Failures*.

Scenario	Component **A**	Component **B**	Component **C**	Component **D**
Initially: Greetings	task	facial expression	limited facial expression	show a big smile
		human body	no human body	try on
During Interaction: User Challenges	verbal abuse	human language	hard to catch	showing difficulties
		tiredness	lack of battery	complain on no time to rest
		got a fever	overheat	cool down
	avoidance	language comprehension	language learning	more efforts in learning human language
		boring language	lack of humor	tell a joke
When Wrong: Handling Failures	color	vision	color sensor	upgrade color sensors
	darkness level	vision	display brightness	end energy saving mode
		vision	display brightness	buy a brighter screen
	sleeve length	vision	camera	buy a better camera
		limbs	do not have limbs	learn by feedback
	style	vision	database	need to lean more knowledge about fashion
		personality	do not have robot of different style	get out of the assembly line to learn about personality
		vision	camera	buy a pair of glasses
	category	classification skills	database	update the commodity labels
		classification skills	classification algorithm	switch to a more advanced classification algorithm
		feet	never wear shoes before	learn by feedback
	shoe upper	fashion concept	database	take a look at some latest fashion magazines

and was initially extracted from *Reddit*[4]. Two researchers independently read through all the punchlines to mark the appropriate self-mockery candidates, and they discussed to conclude 33 punchlines available for the self-mockery construction. They further extended their search to numerous non-academic blogs such as *"Funny Self-Deprecating Quotes and Caption Ideas"*[5] and applied a similar process in selecting potential self-mockery sentences applicable for chatbots. Eventually, 140 potential self-mockery sentences for the chatbot were collected.

However, this method was far from satisfactory. Firstly, researchers generated annotations on a sentence-by-sentence basis, which was time-consuming and prone to bias as people would subjectively perceive humor in the HCI with diverse manners [50]. Secondly, the proportion of the potential self-mockery sentences in the collected humor dataset was merely around 1%, and thus unlikely to provide adequate coverage to the possible cases that the chatbot may face. Besides, according to the observation from the researchers, the collected joke sentences generally mocked human features such as laziness and body shape, which tended to be offensive, and even related to sexual harassment or discrimination. To sum up, most of them cannot be adopted by task-oriented chatbots directly without heavy rephrasing and modification.

[4]https://www.reddit.com/
[5]https://turbofuture.com/internet/Funny-Self-Deprecating-Caption-Ideas

3.2 Generation-Based Method with Designed Template

The generation-based method aimed to compile a new sentence to respond to user input, which is generally more robust than the retrieval-based method if properly designed [54, 63]. To effectively apply this method, we proposed to create templates based on theories of self-mockery in the HHI. According to Ungar et al., self-mockery is to make jokes about oneself based on one's weakness [69]. Then we assumed that mocking on the perceived difference between agents (chatbots, robots, etc.) and humans could be a practical way to achieve the effects of the self-mockery for chatbots. Based on this assumption, we drafted a systematic way of generating self-mockery sentences for chatbot, which was to construct four components first and then fill them into the designed templates for different scenarios (see section 3.3). Specifically, the four components to be applied in the templates would be:

A Task related details
B Human cognitive/physical systems
C Issues in chatbot/machine components
D Human or chatbot solutions

Accordingly, the total pipeline would be: **1)** selecting the specific features and corresponding cognitive systems to mock and generate relative components from **A** to **D**, **2)** adapting and filling these phrases to the pre-defined templates for different scenarios, and **3)** rephrasing the generated sentences to make them more consistent. Figure 1 demonstrates the entire self-mockery generation pipeline.

Table 2: Self-mockery templates and generated examples (refer to Table 5 for Component A selection).

Scenario	Templates	Examples
Initially: Greetings	Welcome! This is <BOT_NAME> from <BUSINESS_NAME>. It's so nice to meet you. <C>, but I can already imagine the smile on your face when <A>!	Welcome! This is <BOT_NAME> from <BUSINESS_NAME>. It's so nice to meet you. **My facial expressions are limited**, but I can already imagine the smile on your face when **you put your new shoes on**!
	Greetings! This is <BOT_NAME> from <BUSINESS_NAME>. I'm your shopping guide who <C>, but I certainly know <A>! How can I help you today?	Greetings! This is <BOT_NAME> from <BUSINESS_NAME>. I'm your shopping guide who **doesn't have human body to put clothes on**, but I certainly know **what suits you best**! How can I help you today?
During Interaction: User Challenge	Please not be that mean. I have been <B>. <D>.	Please not be that mean. I have been **working so long that I think I got a fever**. I really need to **cool down my brain for computation**.
	Sorry, it seems <B>. <D>.	Sorry, it seems **the evolution of human language** is so fast that make the **spinning speed of my processor** hard to catch.
	<SILENCE_BREAKING>. Just to be sure that it is not me that <D>.	Still there? Just to be sure that it is not me that **overwhelmed by human language**.
When Wrong: Handling Failures	Sorry for the wrong <A>, I guess I really need to <D> to improve/upgrade/refresh my <B>.	Sorry for my poor **color vision**. I think it's time to **get a pair of new glasses** to support my **sensors on detecting colors**.
	Sorry for the poor <B>, I find it's urgent to <D> in order to support my <C> (ability/skill).	
	Sorry for the wrong <A>, well let me just <D> for better <C>.	Sorry for the wrong **darkness level**, well let me just **get out of the energy saving mode** for better **display brightness**.
	Sorry for that, I don't have <B> like human so please allow me to <D> to get a concept of it.	
	Sorry for that, I <C> so I need to <D> to get some common knowledge of the <A>.	Sorry for that, I have **never worn shoes before**, so I need to **learn by feedback** to get some common knowledge of **the shoe category**.

Table 1 lists the example components for self-mockery generation with the Generation-Based method.

3.3 Self-Mockery Scenario Construction

For each interaction stage [2], we constructed a scenario suitable to apply self-mockery, and validated them with the design guidelines of the conversational agents to ensure no conflicts [78]. To be specific, three researchers who are experienced in chatbot interaction (all interacted with chatbots at least once a week) derived suitable application scenarios of the self-mockery based on practical purposes of applying self-mockery in the HHI [45]. They discussed to match the purposes to the interaction stages. They further verified the matching with the existing literatures to rigorously structure the self-mockery application scenarios and correspondingly surveyed the potential benefits in the HHI. The selected scenarios are (see section 3.3.1 - 3.3.3 for details):

- *Greetings* [57]: reduce the social distance in ice-breaking [69].
- *User challenge* [43]: facilitate emotional well-being when being cursed [32].
- *Handling failures* [6, 37]: lighten the tense atmosphere when apologizing [45].

The scenarios echo the interaction stages *initially*, *during interaction*, and *when wrong*. Table 2 lists the templates and the generated self-mockery phrases of the three designed scenarios, matching the example components listed in Table 1. In addition, selecting all components from **A** to **D** could support a better understanding of the generated self-mockery, but not all components were required to be explicitly mentioned during the conversation (see Table 3 for details).

Table 3: Frequent components to be explicitly used.

Scenario	A	B	C	D
Initially: Greetings	✓		✓	
During Interaction: User Challenges		✓		✓
When Wrong: Handling Failures	✓	✓	✓	✓

3.3.1 Initially: Greetings. For the *greeting*, previous work pointed out that chatbots with interesting ice-breaking would be more pleasing to users [67]. Expectation management is also an excellent way to improve user experience, especially conducted at the beginning [50, 52]. Hence, based on common greeting sentences in the HHI, we added self-mockery elements derived from mapping

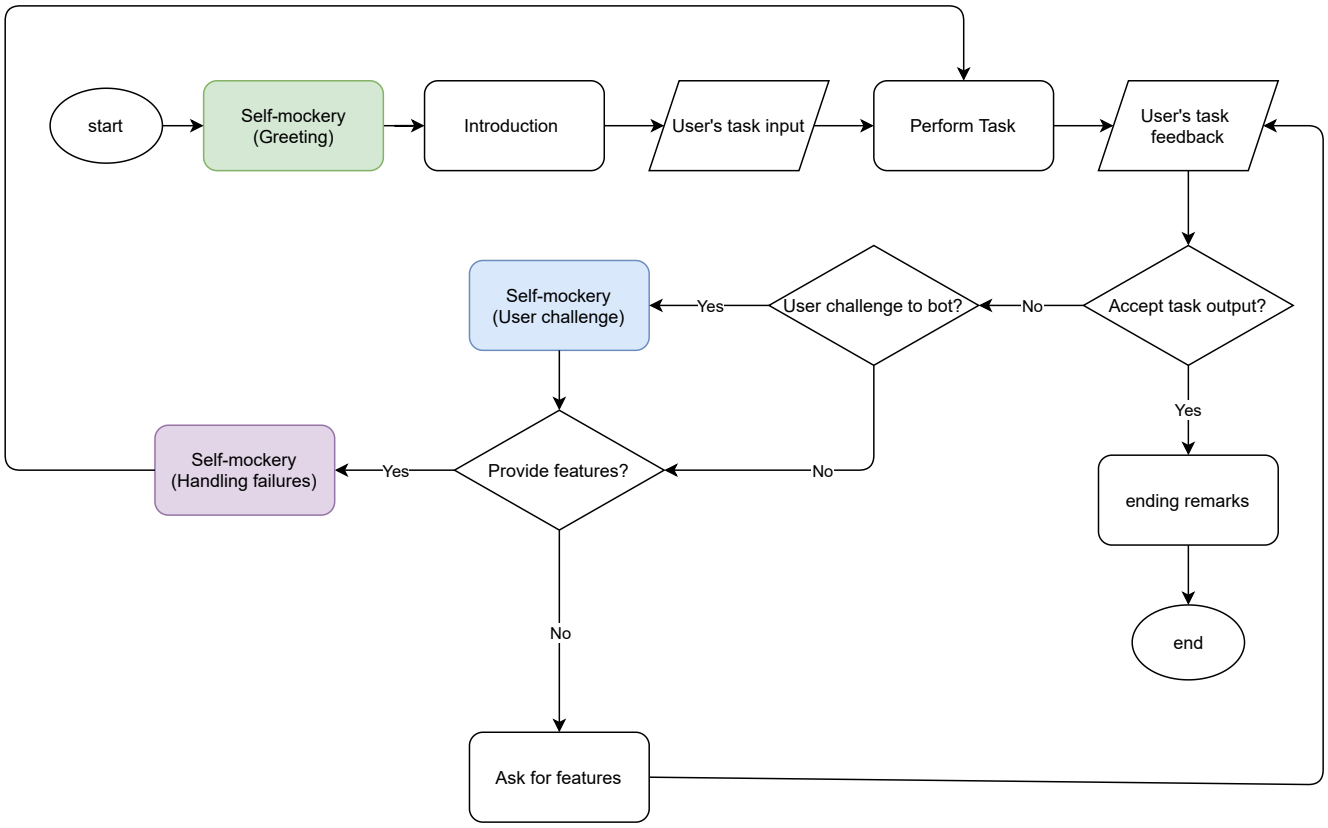

Figure 2: The chatbot interaction logic (with self-mockery).

the chatbot's inability (Component **C**) to task-related background (Component **A**).

3.3.2 During Interaction: User Challenge. Previous work revealed that using submissive characters and tones could make the agents more emotionally intelligent when dealing with the user challenges, i.e., verbal abuse and avoidance [43]. Thus, we applied self-mockery in a submissive way by resembling the human uneasiness (Component **B**) associated with the machine solutions (Component **D**).

3.3.3 When Wrong: Handling Failures. According to the design guidelines [78] for this interaction stage, the chatbots should supply users explanations on why the task could not be completed, which usually incurs task context information. Consequently, we focused on the idea of "make fun of oneself", which means the chatbot would apply self-mockery based on the specific aspect that the chatbot gets wrong. In our simulated online shopping context to evaluate self-mockery, such errors are generally unsuitable recommendations related to product features such as color or style (see section 4.1). Therefore, the self-mockery generation in this scenario would start with the task-related information (Component **A**), i.e., product features, and followed with other components accordingly.

4 CHATBOT IMPLEMENTATION

We implemented our self-mockery chatbot and the baseline chatbot with *RASA*[6], a leading conversational AI platform. To evaluate the self-mockery design on task-oriented chatbots, we simulated an online shopping context where the chatbot served as a shopping guide (gender-neutral), one of the most common real-world applications of the task-oriented chatbots [81]. Integrating such chatbots to conduct customer service is a common approach for the increasingly popular online shopping platforms, and there is a critical need for them to improve user experience [26].

4.1 Product Warehouse for the Chatbot

Inspired by the work of Peng et al. [53], we collected a suit shirt and a pair of sneakers in unisex style from an online shopping platform as our target products. To fully build a virtual "warehouse" for our customer service chatbot, we further collected other relevant product images concerning different product attributes, e.g., style, category, etc. We also used image processing techniques to modify other features, such as color, darkness level, etc. To this end, for each product (suit shirt and shoes), we created 16 alternatives (= 2^4, 4 attributes each containing one target value and one not) to cover all possible product images to show during the interaction when

[6] https://rasa.com/

Table 4: Selected comparison of the baseline chatbot's language and the corresponding self-mockery language.

Scenario	Baseline	Self-mockery
Initially: Greetings	Welcome! This is your shopping guide of the store.	Welcome! This is your shopping guide of the store. It's so nice to meet you. My facial expressions are limited, but I can already imagine the smile on your face when you put on your new shoes!
	Greetings! This is your shopping guide of the store.	Greetings! This is your shopping guide of the store. I don't have human body to put clothes on, but I certainly know what suits you best!
During Interaction: User Challenge	I didn't get that. Could you try again?	Still there? Just to be sure that it is not me that overwhelmed by human language.
	I didn't get that. Could you try to rephrase it?	Oops, the evolution of human language is so fast that makes the spinning speed of my chips hard to catch.
	That's not nice.	Yeah, so I do not have any salary for working all day long, how poor am I!
	I don't know how to respond to that.	Please don't be that mean. I have been working so long that I think I got a fever. I really need to cool down.
When Wrong: Handling Failures	Sorry for the poor eye sight	Sorry for the poor eye sight, I find it's urgent to buy a better camera in order to support my color sensor ability.
	Sorry for the wrong color	Sorry for the wrong color, I guess I really need to buy a pair of glasses to improve my color vision.
	Sorry for the wrong shoe upper	Sorry for the wrong shoe upper, I guess I really need to take a look at the latest shoes product to refresh my fashion concept.
	Sorry for the wrong category	Sorry for the wrong category, I guess I really need to learn more about the commodity labels to improve my classification skills.

Table 5: Simulated product warehouse

Suit Shirt	**Color**	**Darkness**	**Sleeve Length**	**Style**
Target Value	Blue	Dark	Long	Formal
Confusing Value	Red	Light	Short	Informal
Sneaker	**Color**	**Darkness**	**Shoe Upper**	**Category**
Target Value	Blue	Light	High	Sneaker
Confusing Value	Red	Dark	Low	Leather

needed. Details of the product profiles with different attributes can be found in Table 5.

4.2 Interaction Logic with Self-Mockery Chatbot

When users triggered the chatbot, it would send *greetings* containing self-mockery to users and prompt them to describe what they want to buy. The user input was open-ended without any restrictions, and the chatbot would therefore try to infer the user intent and see if it matched the pre-trained model, a feature powered by *RASA*. The chatbot would prompt the users to describe product attributes more accurately if no specific user intention were identified. The chatbot would then display the matched product picture three seconds after the text message was sent to better build user perception towards the chatbot's language. In order to ensure the self-mockery language has gained enough exposure to the users, the chatbot would only identify one most closed attribute that deviates from the users' desired value and correct only this value in the product picture displayed in response to each user input. In this way, the chatbot could present several rounds of different product pictures and send messages containing self-mockery for *handling failures* accordingly before reaching the correct product. When the user performed *user challenge* to the chatbot, such as verbal abuse or avoidance [6, 30], the chatbot would respond with self-mockery designed for this scenario. Considering that some users might not impose the *user challenge* to the chatbot even if they were dissatisfied, we created a **Complain** button that would appear after the chatbot failed to fetch the right product for them, making it more accessible to conduct *user challenge*. After pressing the **Complain** button, a message "Stupid chatbot!" would be automatically sent from the user side, an approach that had been proved to be an indirect but truthful way to measure user perception related to *user challenge* [43]. In addition, to free users from the potential endless loop to interact with the chatbot, we also designed a **Transfer to Human** bottom that appeared after the third failure such that they could opt to click the button to end the conversation by then. Figure 2 demonstrates the logic flow of the chatbot as described above.

4.3 Baseline Chatbot Language

We implemented the baseline chatbot with exactly the same interactive logic except for the self-mockery language to make the evaluation fair. Four researchers tested several real world chatbots, such as those on *Facebook* and *Amazon*, to gain experience on how existing chatbots interacted with users without self-mockery in different scenarios. They further discussed and concluded the language of the baseline chatbot in different scenarios, namely *greetings*, *user challenge*, and *handling failures*. Table 4 illustrates the comparison of the baseline and the self-mockery chatbot languages.

In scenario *greetings*, the baseline chatbot only provided concise information and brief openings, similar to most of the chatbots we

tested, while our self-mockery robot started by kicking off with self-mockery. For the *user challenge* scenario, the baseline chatbot either showed a sense of confusion or refused to respond to the abusive behaviors of users. In fact, only a few of the tested commercial chatbots directly pointed out that the users were not with appropriate behaviors or used engaging words to ask the users to rephrase. In contrast, the self-mockery chatbot would try to exert self-mockery to smooth the atmosphere. For the *handling failures* scenario, the baseline chatbot generally apologized after making mistakes but presented no signs of remedy or further instructions, while the self-mockery chatbot would mock itself on relevant issues (see section 3.3.3). In summary, to make it representative in general cases, we designed the language of the baseline chatbot to incorporate as many of the essential features of the commercial chatbots as possible.

5 EVALUATION

To answer **RQ1** and **RQ2**, we conducted the hypothesis testings. To further address **RQ2**, we compared the perceived helpfulness of adapting self-mockery for the chatbot in different scenarios towards individual characteristics of social intelligence. Finally, we conducted regression analyses to understand the impact of individual factors on their perception towards the effectiveness of incorporating self-mockery on the chatbot's social intelligence for **RQ3**.

5.1 Hypotheses and Measurements

In the HHI, self-mockery is a kind of humor which helps to increase the positive emotions between people [17]. To trigger such effects, self-mockery language should be funny enough [45]. If done appropriately, it can also enhance the emotional solidarity and sense of identity with others [69]. Hence, successful self-mockery language can improve social intelligence in the HHI, which is defined as the ability of an individual to perform appropriate social behaviors in order to achieve expected goals [9]. Therefore, to evaluate the self-mockery language for the task-oriented chatbot, we hypothesized that (all measured in standard 7-point Likert scale, where 1 for the most negative impression and 7 for the most positive):

- **H1** Compared to the baseline chatbot, language used by the self-mockery chatbot is significantly funnier [45] (*H1a*), but appropriateness [69] does not have a significant difference. (*H1b*).
- **H2** Compared to the baseline chatbot, users' overall satisfaction level [58] is significantly improved such that they were more satisfied with the self-mockery chatbot (*H2a*) and inclined to use the self-mockery chatbot again (*H2b*).
- **H3** Compared to the baseline chatbot, all measured characteristics of the social intelligence [13], to be specific, *damage control* (*H3a*), *thoroughness* (*H3b*), *manners* (*H3c*), *moral agency* (*H3d*), and *emotional intelligence* (*H3e*), are significantly improved in the self-mockery chatbot.

To sum up, the hypothesis testings on users' perceptions towards the languages of (**H1**), the satisfaction with interactions with (**H2**), and the social intelligence of (**H3**) are conducted on chatbots with/without self-mockery.

5.2 Participant and Procedure

With the approval of our institution's IRB, we recruited 28 participants (10 **F**emale, 15 **M**ale, 3 **N**ot **A**vailable; Age: $Mean = 21.07, S.D. = 1.49$) through online advertising, social media, and word-of-mouth from a local university. The participants were relatively young population as this is the majority group of people who frequently shop online and use chatbots [31, 80, 81]. All of them were familiar with online shopping and reported that they had previous experience interacting with customer service chatbots on online shopping platforms.

After the participants consented to the study, we introduced the experiment procedure and asked them to fill out the questionnaire collecting relevant background information. We also collected the self-reported rating (on a 7-point Likert scale) of the participants towards the three relevant individual factors that reflect their individual preferences and behaviors towards the task-oriented chatbots (see section 6.4) [6]. The participants were later informed that they needed to complete two online shopping tasks and attended a brief interview to receive compensation for participating in the study. They would interact with the baseline chatbot and the self-mockery chatbot in two separate sessions to buy two different target products (see section 4). The entire study was conducted online, and participants were interacted freely with the chatbots to simulate usual use cases and communicated with researchers only if they encountered technical issues. Specifically, we privately counterbalanced the study with Latin square design to minimize the order effect and context effect:

- Suit shirt (baseline) - Sneakers (self-mockery)
- Sneakers (baseline) - Suit shirt (self-mockery)
- Sneakers (self-mockery) - Suit shirt (baseline)
- Suit shirt (self-mockery) - Sneakers (baseline)

In each task, the participants were firstly shown with the corresponding photo of the target product to buy (see section 4.1) and then received the link to the customer service chatbot, which served as a shopping guide. The participants would then start to interact with the chatbot by describing what product they needed to buy. They could also opt to end the interaction if they did not want to interact with the chatbot anymore. To thoroughly test the efficacy of the chatbot self-mockery design on social intelligence, we addressed users that they should interact freely with the chatbot in their usual manners, which could involve any behaviors of *user challenge* such as verbal abuse and avoidance [43].

Finally, after the participants interacted with the two chatbots in counterbalanced order and correspondingly answered the questionnaire, they were then informed about the self-mockery design in different scenarios. They would be asked to give the ratings on the helpfulness of self-mockery design to the five measured characteristics of social intelligence [13] based on the three scenarios respectively. Before the end of the study, we conducted a semi-structured interview with the participants asking about their general perception towards the self-mockery chatbot, advantages & disadvantages of using self-mockery to improve the specific characteristic of social intelligence (in specific scenarios), and general comments on the design of the self-mockery language. We applied thematic analysis [10] to the interview transcripts to understand participants' perceptions of the chatbot's self-mockery language.

The study generally lasted around 30-50 minutes, and each participant received $6 for attentive participation.

6 RESULTS

Wilcoxon signed-rank test [75] was performed to assess the difference in the participants' ratings on measurements mentioned in section 5.1. The test also affirmed that the context difference, ordering, and gender distribution did not significantly influence the quantitative results. Table 6 summarizes the overall quantitative results of **RQ1** and **RQ2**. Moreover, in this section, we use *[Participant ID, Gender, Age]* to represent the participants to support the presentation of the qualitative findings.

6.1 Manipulation Check

The manipulation check for the chatbot with self-mockery design showed that the manipulation is effective ($W = 43.00, p = 7.00 \times 10^{-4}$). The self-mockery chatbot was perceived to use self-oriented humor ($Mean = 4.83, S.D. = 1.43$) compared to the baseline chatbot ($Mean = 2.88, S.D. = 1.74$).

6.2 Perception on Self-Mockery Chatbot (RQ1)

6.2.1 Evaluation on the Language. We asked participants to evaluate if the chatbot's language was funny after they interacted with each of the chatbots and found that the participants perceived the language of self-mockery chatbot ($Mean = 4.46, S.D. = 1.53, W = 38.00, p = 5.67 \times 10^{-3}$) was significantly funnier than the baseline chatbot ($Mean = 3.00, S.D. = 1.85, W = 38.00, p = 5.67 \times 10^{-3}$); *H1a* accepted. In addition, participants perceived no significant difference between the appropriateness of languages employed by the baseline ($Mean = 5.33, S.D. = 1.37, W = 146.00, p = 7.39 \times 10^{-1}$) and the self-mockery chatbots ($Mean = 5.00, S.D. = 1.63, W = 146.00, p = 7.39 \times 10^{-1}$); *H1b* accepted.

Although most participants considered the current design of the self-mockery language for chatbots as funny and appropriate, they gave insightful feedback on how to improve the language further. For example, [P1, F, 21] and [P16, M, 20] suggested that the self-mockery can be shorter and P3 considered using homonyms jokes is a good way to make the self-mockery concise. [P27, NA, 24] proposed that self-mockery used for the *greeting* could be more in "chit-chat" style, e.g., combing self-mockery with the weather as the opening. Moreover, three participants suggested that the frequency of applying self-mockery should be controlled; otherwise, the chatbot would be regarded as *"glib"* ([P1, F, 21]), and the self-mockery would become *"not interesting anymore"* ([P14, M, 21]). Previous work in the HHI also found that self-mockery would be ineffectual if used too frequently [69].

6.2.2 User Satisfaction. Compared to the baseline chatbot ($Mean = 4.08, S.D. = 1.55, W = 27.50, p = 5.64 \times 10^{-4}$), participants were significantly more satisfied with the interaction of the self-mockery chatbot ($Mean = 5.54, S.D. = 1.22, W = 27.50, p = 5.64 \times 10^{-4}$); *H2a* accepted. In addition, they were also significantly more preferred to use the self-mockery chatbot again ($Mean = 5.92, S.D. = 1.08, W = 28.00, p = 3.78 \times 10^{-4}$) than the baseline chatbot ($Mean = 3.96, S.D. = 1.72, W = 28.00, p = 3.78 \times 10^{-4}$); *H2b* accepted.

In the post-study interview, the majority of the participants (19/28) expressed that equipping chatbots with self-mockery was an effective way to improve the user experience, making chatbots more human-like and increasing users' willingness to interact. Specifically, [P4, F, 21] was impressed that self-mockery chatbot *"gracefully reacted to malicious language"*, and [P16, M, 20] pointed out that self-mockery was effective as *"customers would feel harder to complain to a smiling face"*. The concerns of using self-mockery were mainly on the frequency and user patience. [P23, M, 22] addressed that *"showing off humor consciously is not a good strategy"*, and [P19, M, 22] added that algorithms behind the chatbot were more vital to users' tolerance as *"the extent of failure that self-mockery could cover is limited"*.

6.3 Perception on the Chatbot's Social Intelligence (RQ2)

6.3.1 Overall Perception Regarding Different Social Intelligence Characteristics. Regarding the perceived characteristics of social intelligence, we measured all of them in the questionnaire after the participants interacted with the baseline or the self-mockery chatbot, except *personalization* due to context constrain (see section 4). The participants perceived the self-mockery chatbot exhibited significantly higher social intelligence in *damage control* (*H3a* accepted) and *emotional intelligence* (*H3e* accepted), while remained comparable performance in *thoroughness* (*H3b* rejected), *manners* (*H3c* rejected), and *moral agency* (*H3d* rejected). Overall, **H3** is only partially accepted (2/5), and we received comments parallel to the above findings in the post-study interviews.

Damage Control. Participants commented that the self-mockery chatbot could make them less angry when making mistakes ([P8, F, 23], [P27, NA, 24]) and increase their tolerance on chatbot's failure ([P9, F, 20]). Specifically, [P3, M, 19] highlighted that self-mockery is *"a good way for chatbots to relive the embarrassment during the conversation"*, an effect similar to self-mockery used in the HHI [45]. [P27, NA, 24] followed that although overall speaking self-mockery could improve the chatbot's social intelligence on *damage control*, the effectiveness would be varied for different scenarios: in the scenario *when wrong*, self-mockery language could not help in any sense as he wanted to get the right product as soon as possible; in the scenario *user challenge*, this approach would be very effective as it largely appeased his anger.

Thoroughness. Considering the self-mockery chatbot preserved a similar language pattern to the baseline chatbot, it is not surprising that both chatbots were perceived with a high level of *preciseness in using the language* [48], the definition of *thoroughness*. Four participants ([P3, M, 19], [P7, M, 24], [P8, F, 23], [P21, NA, 20]) explicitly mentioned that the self-mockery and the baseline chatbot were consistent in different styles; e.g., according to [P3, M, 19], the self-mockery chatbot gave people an encaustic and perhaps a little bit talkative impression, while the baseline chatbot seemed more like an introverted but professional shopping guide.

Manners. Most participants (20/28) considered both self-mockery and baseline chatbot had polite behavior and conversation habits since they both exhibited the sorry attitude after making mistakes. For [P24, M, 21], both chatbots were *"polite since they always apologize to me when they found themselves prompted me with the wrong*

Table 6: The quantitative results of participants' evaluation with the baseline and the self-mockery chatbot, where the p-values (-: p > .100, +: .050 < p < .100, *:p < .050, **:p < .010, ***:p < .001) are reported.

RQ	Measurements	Baseline Mean/S.D.	Self-mockery Mean/S.D.	Statistics		Hypotheses
				W	p-value	
MC	Manipulation Check	2.88/1.74	4.83/1.43	43.00	0.001^{***}	
RQ1	Funniness	3.00/1.85	4.46/1.53	38.00	0.005^{**}	H1a accepted
	Appropriateness	5.33/1.37	5.00/1.63	146.00	0.739^{-}	H1b accepted
	Satisfaction	4.08/1.55	5.54/1.22	27.50	0.001^{***}	H2a accepted
	Use Again	3.96/1.72	5.92/1.08	28.00	0.000^{***}	H2b accepted
RQ2	Damage Control	4.13/1.79	5.29/1.49	40.50	0.008^{**}	H3a accepted
	Thoroughness	5.38/1.32	5.42/1.35	112.00	0.450^{-}	H3b rejected
	Manners	5.00/1.38	4.88/1.51	100.00	0.581^{-}	H3c rejected
	Moral Agency	4.83/1.57	4.79/1.66	140.50	0.531^{-}	H3d rejected
	Emotional Intelligence	3.42/1.71	5.38/1.28	15.00	0.000^{***}	H3e accepted

products". [P10, F, 21] simply regarded self-mockery as an alternative yet appropriate way of service language; as for her, *"demonstrating consideration professionally"* is fundamental to judge *manners*. Additionally, five participants were quite divided on the efficacy of using self-mockery design to improve *manners* of the chatbot. [P26, M, 22] addressed that *"self-mockery chatbot can light up the tense atmosphere when the users are angry, a smart way to keep its dignity"*, while [P5, M, 21] insisted that the use of self-mockery gave him a feeling of non-seriousness in the task and thus unprofessional.

Moral Agency. As for *moral agency*, the majority of the participants (22/28) gave similar scores (same or neighboring choices) to both chatbots since both could understand their input and converse properly without violating any social conventions. [P11, M, 20] said that it is natural for people to make mistakes, and chatbots could exhibit a high level of *moral agency* simply by *"realizing and acknowledging the mistakes"*, a common human virtue seemed irrelevant to use self-mockery or not. Yet, two participants held different ideas. For example, [P26, M, 22] pointed out that using self-mockery, especially during *user challenge*, could reduce hostility and somehow educate people to be polite.

Emotional Intelligence. Participants commented that when interacting with the self-mockery chatbot, besides reducing the anger, the chatbot was able to create a more relaxing atmosphere ([P9, F, 20], [P14, M, 21], [P17, F, 21]). [P18, M, 22] followed that apart from adjusting the tense atmosphere during the conversation, the self-mockery design *"powered the chatbot's capability of emotional expression and personification"*. [P7, M, 24] addressed that he really appreciated that chatbot tried to use self-mockery not to let users down. Such behavior made it *"emotionally like human a lot"*, especially during *user challenge*.

6.3.2 Perceived Helpfulness in Different Scenarios. In order to better understand the effects of the self-mockery design on the various characteristics of the chatbot's social intelligence under different scenarios, we further collected the user ratings on a 7-point Likert scale (1 for the least helpful, 7 for the most) on the helpfulness of the self-mockery design in each scenario towards each characteristic. The result showed no significant difference in the rating of the self-mockery design in three scenarios towards *thoroughness* and *manners*. While weak significance existed in *damage control* and *emotional intelligence*, there was strong significance in *moral agency*. Friedman's test [56] showed that there are significant differences among the three scenarios regarding *moral agency*. Further testing by Wilcoxon signed-rank test [75] showed that the ratings in scenario *user challenge* and scenario *when wrong* were significantly higher than *greetings*. This means that participants considered that for *moral agency*, self-mockery design in scenario *greetings* was significantly less helpful compared to the other two scenarios. In summary, although the quantitative analysis of the detailed helpfulness scores rated by participants did not yield a conclusive result, the trend is that the self-mockery strategy of chatbots is likely to be the most beneficial when designed to handle *user challenge* and least helpful for *greetings*. Detailed results are showed in Table 7 and Figure 3.

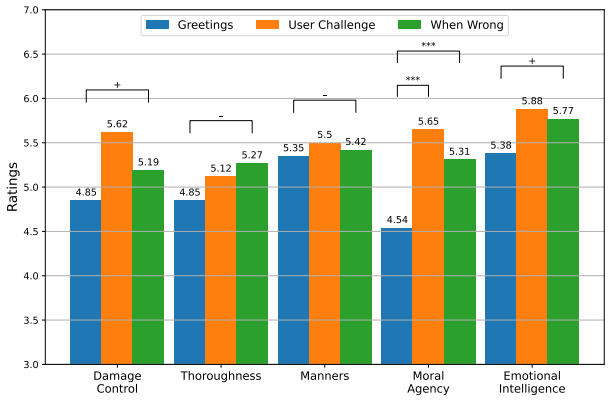

Figure 3: Means of the perceived helpfulness of self-mockery to social intelligence, where the p-values (-: p > .100, +: .050 < p < .100, *:p < .050, **:p < .010, ***:p < .001) is reported.

6.4 Impact of Individual Factors (RQ3)

In this section, we further explore whether individual factors of the participants (social orientation towards chatbots [39, 40], service orientation [37], and experience with chatbots [6]) can influence

Table 7: Comparison of the perceived helpfulness of self-mockery to social intelligence in the three scenarios, where the p-values (-: p > .100, +: .050 < p < .100, *:p < .050, **:p < .010, ***:p < .001) are reported.

	Greetings	User Challenge	When Wrong	Statistics	
	Mean/S.D.	Mean/S.D.	Mean/S.D.	W	p-value
Damage Control	4.85/1.46	5.62/1.36	5.19/1.36	5.56	0.062^{+}
Thoroughness	3.79/1.41	3.79/1.41	5.17/1.03	13.50	0.607^{-}
Manners	3.58/1.32	3.58/1.32	4.92/1.04	9.00	0.405^{-}
Moral Agency	3.04/1.27	3.04/1.27	5.38/1.11	13.50	0.000^{***}
Emotional Intelligence	3.33/0.99	3.33/0.99	5.46/1.00	4.00	0.073^{+}

their perceptions of self-mockery language of the chatbot as either Moderating Variables (MVs) or Independent Variables (IVs). To accomplish this, we standardized IVs with Z-scores accordingly and conducted two sets of regression analyses. Additionally, we calculated the correlations between each pair of IVs and the Variance Inflation Factor (VIF) score [18] of each variable in the regression models, the result of which indicated that multicollinearity did not occur in our models.

6.4.1 Individual Factors as Moderator Variables. We firstly applied a set of hierarchical regression models [12, 55] to explore whether the individual factors had a moderating effect on the relation between self-mockery language design and the measured Dependent Variables (DVs), including language evaluation (*funniness, appropriateness*), user satisfaction (*satisfaction, use again*), and overall perception regarding different social intelligence characteristics (*damage control, thoroughness, manners, moral agency, emotional intelligence*). All nine models ran for the DVs have self-mockery language design as their IV (0 represents baseline chatbots, 1 represents self-mockery chatbots) and three standardized participants' individual factors as MVs. We found no significant difference between the model with interaction terms (between IVs and MVs) and the model without those terms. This fact suggested that all the measured individual factors of the participants did not moderate the relationship between whether applying self-mockery design on the chatbot and participants' perceptions of the chatbot's language, overall satisfaction, and any characteristic of the chatbot's social intelligence.

6.4.2 Individual Factors as Independent Variables. We then investigated how individual factors affect the perception of self-mockery design's helpfulness towards five different characteristics of social intelligence in three different scenarios (total 15 DVs in this case). As a result, we constructed fifteen linear regression models, each of which predicted one DV by incorporating the three user factors as IVs. We focused on statistically significant results.

Social Orientation toward Chatbots. Social orientation measures the users' preference for human-like social interactions with agent interfaces, indicating whether users like to treat the agent as sociable actors or not [6, 39, 40]. In our study, participants with higher social orientation were more likely to perceive self-mockery language as helpful in regard to all characteristics of social intelligence under all scenarios, either significantly or marginally (with p-value < 0.1 and $\beta > 0$ for all models). These quantitative results indicate that people with a higher social orientation towards chatbots are more likely to view self-mockery language as a natural interaction with humans rather than a mechanical response, making them more inclined to appreciate self-mockery design as helpful. Table 8 summarizes the quantitative results of this part.

Service Orientation. In contrast to people with a strong relational orientation, Lee et al. found that people with a strong utilitarian orientation were more concerned with efficiency and accuracy of service than the actual interaction [37]. Nevertheless, we did not find any statistical differences across users with different service orientations (with $p > 0.05$ for all models), suggesting service orientation did not significantly affect the perceived helpfulness of self-mockery language design.

Experience with Chatbots. Participants' self-reported prior experience with chatbots is used to determine their familiarity with the chatbot interaction [6]. The result indicated that participants with more experience with chatbots were significantly more likely to have a more positive perception of the helpfulness of self-mockery language regarding chatbots' *thoroughness* in *user challenge* ($\beta = 0.413, p = 0.021$). However, since this is the only significant result among all fifteen models, thus it is highly possible that the experience with chatbots only slightly affects users' perception of the helpfulness of the self-mockery design towards chatbots' social intelligence in general.

7 DISCUSSION

The self-mockery language generated with our designed pipeline was perceived as appropriate yet funny by the participants, and applying it also improved their overall satisfaction towards the chatbot. Incorporating self-mockery into the task-orientated chatbot was also demonstrated to raise participants' perception of its social intelligence partially. Further study indicated that participants with a higher social orientation towards the chatbot would perceive the self-mockery more helpful on all characteristics of chatbots' social intelligence in all designed scenarios, while other measured individual factors did not appear to impact participants' perception. In this section, we discussed theoretical reflections, derived design considerations to apply self-mockery to task-orientated chatbots, generalized the self-mockery design pipeline to more diversified presentations, and addressed the limitations to be explored in the future.

7.1 Theoretical Reflections

As one of our research goals, we proposed self-mockery generation method for task-orientated chatbots to improve their social

Table 8: Standardized coefficients (β) and p-values (-: p > .100, +: .050 < p < .100, *:p < .050, **:p < .010, ***:p < .001) of the Social Orientation toward Chatbots (IV) in the regression models.

	Greetings		User Challenge		When Wrong	
	β	p-value	β	p-value	β	p-value
Damage Control	0.346	0.081^{+}	0.352	0.076^{+}	0.583	0.001^{**}
Thoroughness	0.620	0.001^{**}	0.473	0.009^{**}	0.529	0.006^{**}
Manners	0.513	0.008^{**}	0.408	0.037^{*}	0.352	0.070^{+}
Moral Agency	0.441	0.019^{*}	0.636	0.000^{***}	0.355	0.067^{+}
Emotional Intelligence	0.444	0.013^{*}	0.362	0.059^{+}	0.338	0.092^{+}

intelligence (see section 3). However, it is unclear that whether the self-mockery applied by the chatbots can be as robust as that applied by humans. From the HHI perspective, the coverage of the self-mockery applied in our study is a subset of all categories that used in the HHI, according to the previous survey [45]. For example, self-mockery can be used for counterattack and is particularly useful in debate.

Moreover, previous research demonstrated that, similar to the HHI, agents with self-mockery are more likely to receive support from the users [11]. In the HHI, self-mockery is usually regarded as an indicator of shyness and low self-esteem; therefore, when providing support to self-mockery users, people generally feel more capable and confident, making them more inclined to do so [24, 34]. This fact might be another implicit factor in adopting self-mockery, an effective strategy for chatbots to improve their social intelligence.

7.2 Design Considerations

In this part, we discussed key design considerations, including the usage pattern of self-mockery from the angle of managing user expectation, whether to apply self-mockery based on the chatbot's designed persona, and when to appropriately use self-mockery during the conversational flow.

7.2.1 Manage User Expectation. Previous studies suggested that using humor could raise unrealistically high expectations of the users on the agents' capabilities as humor is often regarded as a crucial social feature in the HHI [16, 41]. However, applying self-mockery – self-oriented humor – might give agents a chance to display their potential vulnerabilities [11], which could regulate users' expectations to some extent. Still, it is worth noting that overusing self-mockery could induce a counterproductive effect (see section 6.2.1). Therefore, designers may consider using limited self-mockery for handling critical events or displaying it occasionally in an "Easter Egg" style to add to people's delight [39]. In addition, the design of self-mockery can be more adaptive, e.g., combining with recent news ([P13, F, 20]), to give people a refreshing impression of the chatbot's social intelligence when viewed as a social actor [59].

7.2.2 Maintain a Consistent Persona. Prior research showed that maintaining a consistent persona is fundamental for the perceived naturalness of a conversational agent [33]. In fact, the consistency of the persona is vital to chatbots' perceived social intelligence, especially on *thoroughness* (see section 6.3.1). Therefore, designers need to determine the chatbot's persona based on the application contexts and check if using self-mockery is compatible with the defined persona before employing this rhetorical device. For example, the self-mockery may not be appropriate for the chatbots designed to be AI Doctors [72], as humorous utterances and professional advice by the chatbots as physicians might leave users with a divided impression and further damage their trust in the chatbots [13].

7.2.3 Align with the Conversation. Apart from validating whether self-mockery is compatible with the chatbot's design persona, when to pop up self-mockery should also be carefully considered to suit the conversational flow, especially for task-orientated chatbots [78]. As humor by the chatbots is sometimes regarded as non-task-related talk [19, 70], popping up self-mockery might not be appropriate when users are aiming to complete tasks of specific goals [21]. For instance, [P19, M, 22] mentioned that he would be uncomfortable if a chatbot for mobile banking talked to him about his investment with self-mockery, but it would probably be a good option to apply self-mockery at the opening for *greetings*. As during the conversation, he would only want to acquire accurate and comprehensive information and the self-mockery appeared during this conversation flow would make him feel that the bank was not taking his financial information seriously. To conclude, we suggested that the self-mockery design for a chatbot requires a good understanding of the conversation flow such that the self-mockery language could pop up at the right timing.

7.3 Facilitating Generation of Diversified Self-Mockery Presentations

In this work, we only focused on the text-based self-mockery for task-orientated chatbots. In fact, users can perceive multimedia-based humor better than pure text-based humor [47]. For example, the self-mockery designed with our proposed pipeline can be further forged into humorous "Memes", one of the most shared contents on various social networks [20]. The design of such memes can be inspired by the selected components of the self-mockery, especially Component **B** & **C**, i.e., raw images of corresponding human cognitive systems and machine components (see section 3.2). The emoji can also be applied to enrich the self-mockery language of the chatbots, e.g., using an apologetic emoji face when handling failures with self-mockery to ensure the apology is properly expressed with humorous utterance.

Apart from the chatbots, agents such as robots [53] are also frequently used by people in their daily lives. Previous work demonstrated that robots are able to produce self-mockery laughter that can be perceived by humans [46]. With our proposed pipeline, the robots' self-mockery can be further enhanced by presenting gestures; e.g., when robots fail to select the product in the required

style (Component **A**), they can therefore pointing to their eyes (Component **C** related concepts) and making self-mockery on the vision problem (see Table 1).

7.4 Limitations

Our work has several limitations. First, we only tested the self-mockery chatbot with relatively younger user groups. The previous study indicates that, different from younger users, older users focus more on the chatbots' efficiency and effectiveness instead of their social skills [23]. Still, determining whether to apply self-mockery for task-orientated chatbots designed for older people would need further evaluation on their perception towards it. Second, although we proposed a systematic pipeline to generate self-mockery sentences for a chatbot, this process is not fully automated. It requires selecting the key components to compile the self-mockery and tailoring the sentence after filling the templates (see section 3.2). Future work could apply relevant NLP techniques such as commonsense networks [66] and transformer models [74] to assist humans in components selection and sentence tailoring. Finally, the participants only interacted with the chatbot as "one-time" customers, as this is how the majority of the chatbots are developed in real-world applications [13]. However, there is a trend that more and more chatbots start to have "memories" and interact with people *over time* [42]. Future work could be done on updating the self-mockery design from "one-time" to *over time*.

8 CONCLUSION

In this paper, we designed a pipeline to generate self-mockery language for the task-oriented chatbots based on the previous research of self-mockery in the HHI. We applied the designed self-mockery generation pipeline to a customer service context and consequently implemented a shopping guide chatbot with self-mockery expression. The results showed that the self-mockery language of the chatbot was perceived as funny yet appropriate and significantly improved users' overall satisfaction during the interaction with the chatbot. In terms of the characteristics of the chatbot's social intelligence, the self-mockery design significantly improved *damage control* and *emotional intelligence* with comparable performance in other measured characteristics. Analysis of the participants' self-reported individual factors suggested that those with a higher social orientation towards chatbots [39, 40] would better appreciate the helpfulness of self-mockery design on chatbot's overall social intelligence in all scenarios. In contrast, other measured factors, e.g., service orientation [37] and familiarity with the chatbot technology [6], were only observed with negligible impact in this case. Finally, we further concluded this paper with design considerations of self-mockery on the task-oriented chatbot and possible generalizability of the proposed self-mockery generation pipeline to accommodate a more diversified form of self-mockery presentation. Future work could be done in enhancing self-mockery design with suitable NLP tools and testing the self-mockery of task-oriented chatbots with more diverse user groups. Further explorations could be focused on applying self-mockery to agents that are designed to treat social impairment, considering that presenting self-mockery humor could make users feel more confident [11]. Ideally, if the chatbots' self-mockery skills become mature enough, they might be able to train users on the skill of humor or other more complex tasks.

ACKNOWLEDGMENTS

Many thanks to the anonymous reviewers for their insightful suggestions. We thank Yiduo Yu for his valuable inputs. This work is supported by the Hong Kong General Research Fund (GRF) with grant No. 16204819.

REFERENCES

[1] Eleni Adamopoulou and Lefteris Moussiades. 2020. Chatbots: History, technology, and applications. *Machine Learning with Applications* 2 (2020), 100006. https://doi.org/10.1016/j.mlwa.2020.100006

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. *Guidelines for Human-AI Interaction.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[3] Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the The 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. 29–41.

[4] Antonella De Angeli and Sheryl Brahnam. 2008. I hate you! Disinhibition with virtual partners. *Interacting with Computers* 20, 3 (2008), 302–310. https://doi.org/10.1016/j.intcom.2008.02.004 Special Issue: On the Abuse and Misuse of Social Agents.

[5] Theo Araujo. 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 (2018), 183–189. https://doi.org/10.1016/j.chb.2018.03.051

[6] Zahra Ashktorab, Mohit Jain, Q. Vera Liao, and Justin D. Weisz. 2019. *Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300484

[7] Jaime Banks. 2019. A perceived moral agency scale: Development and validation of a metric for humans and social machines. *Computers in Human Behavior* 90 (2019), 363–371. https://doi.org/10.1016/j.chb.2018.08.028

[8] Lucile Bechade, Guillaume Dubuisson Duplessis, and Laurence Devillers. 2016. Empirical study of humor support in social human-robot interaction. In *International Conference on Distributed, Ambient, and Pervasive Interactions*. Springer, 305–316. https://doi.org/10.1007/978-3-319-39862-4_28

[9] Kaj Björkqvist, Karin Österman, and Ari Kaukiainen. 2000. Social intelligence-empathy= aggression? *Aggression and violent behavior* 5, 2 (2000), 191–200. https://doi.org/10.1016/S1359-1789(98)00029-9

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[11] Jessy Ceha, Ken Jen Lee, Elizabeth Nilsen, Joslin Goh, and Edith Law. 2021. Can a Humorous Conversational Agent Enhance Learning Experience and Outcomes?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 685, 14 pages. https://doi.org/10.1145/3411764.3445068

[12] Joseph E Champoux and William S Peters. 1987. Form, effect size and power in moderated regression analysis. *Journal of Occupational Psychology* 60, 3 (1987), 243–255. https://doi.org/10.1111/j.2044-8325.1987.tb00257.x

[13] Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction* 37, 8 (2021), 729–758. https://doi.org/10.1080/10447318.2020.1841438

[14] Hyojin Chin, Lebogang Wame Molefi, and Mun Yong Yi. 2020. *Empathy Is All You Need: How a Conversational Agent Should Respond to Verbal Abuse.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376461

[15] Hyojin Chin and Mun Yong Yi. 2019. Should an Agent Be Ignoring It? A Study of Verbal Abuse Types and Conversational Agents' Response Styles. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312826

[16] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland

Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705

[17] Matthew J Cousineau. 2016. Accomplishing Profession through Self-Mockery. *Symbolic Interaction* 39, 2 (2016), 213–228. https://doi.org/10.1002/symb.217

[18] Trevor A Craney and James G Surles. 2002. Model-dependent variance inflation factor cutoff values. *Quality engineering* 14, 3 (2002), 391–403. https://doi.org/10.1081/QEN-120001878

[19] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[20] Abhimanyu Dubey, Esteban Moro, Manuel Cebrian, and Iyad Rahwan. 2018. MemeSequencer: Sparse Matching for Embedding Image Macros. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1225–1235. https://doi.org/10.1145/3178876.3186021

[21] Daniëlle Duijst. 2017. Can we improve the user experience of chatbots with personalisation. *Master's thesis. University of Amsterdam* (2017). https://doi.org/10.13140/RG.2.2.36112.92165

[22] Haiyan Fan and Marshall Scott Poole. 2006. What is personalization? Perspectives on the design and implementation of personalization in information systems. *Journal of Organizational Computing and Electronic Commerce* 16, 3-4 (2006), 179–202. https://doi.org/10.1080/10919392.2006.9681199

[23] Asbjørn Følstad and Petter Bae Brandtzaeg. 2020. Users' experiences with chatbots: findings from a questionnaire study. *Quality and User Experience* 5, 1 (2020), 1–14. https://doi.org/10.1007/s41233-020-00033-2

[24] William P Hampes. 2006. Humor and shyness: The relation between humor styles and shyness. (2006). https://doi.org/10.1515/HUMOR.2006.009

[25] Michael Haugh. 2011. Humour, face and im/politeness in getting acquainted. *Situated politeness* (2011), 165–184.

[26] Miri Heo, Kyoung Jun Lee, et al. 2018. Chatbot as a new business communication tool: The case of naver talktalk. *Business Communication Research and Practice* 1, 1 (2018), 41–45. https://doi.org/10.22682/bcrp.2018.1.1.41

[27] Nabil Hossain, John Krumm, Tanvir Sajed, and Henry Kautz. 2020. Stimulating creativity with funlines: A case study of humor generation in headlines. *arXiv preprint arXiv:2002.02031* (2020). https://doi.org/10.48550/arXiv.2002.02031

[28] Shafquat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. 2019. A survey on conversational agents/chatbots classification and design techniques. In *Workshops of the International Conference on Advanced Information Networking and Applications*. Springer, 946–956. https://doi.org/10.1007/978-3-030-15035-8_93

[29] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N. Patel. 2018. *Convey: Exploring the Use of a Context View for Chatbots*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3173574.3174042

[30] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 895–906. https://doi.org/10.1145/3196709.3196735

[31] Mohammad Hossein Moshref Javadi, Hossein Rezaei Dolatabadi, Mojtaba Nourbakhsh, Amir Poursaeedi, and Ahmad Reza Asadollahi. 2012. An analysis of factors affecting on online shopping behavior of consumers. *International journal of marketing studies* 4, 5 (2012), 81. https://doi.org/10.5539/IJMS.V4N5P81

[32] Samuel Juni and Bernard Katz. 2001. Self-effacing wit as a response to oppression: Dynamics in ethnic humor. *The Journal of general psychology* 128, 2 (2001), 119–142. https://doi.org/10.1080/00221300109598903

[33] Yelim Kim, Mohi Reza, Joanna McGrenere, and Dongwook Yoon. 2021. Designers Characterize Naturalness in Voice User Interfaces: Their Goals, Practices, and Challenges. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 242, 13 pages. https://doi.org/10.1145/3411764.3445579

[34] Nicholas A Kuiper and Nicola McHale. 2009. Humor styles as mediators between self-evaluative standards and psychological well-being. *The Journal of psychology* 143, 4 (2009), 359–376. https://doi.org/10.3200/JRLP.143.4.359-376

[35] Rohit Kumar, Hua Ai, Jack L Beuth, and Carolyn P Rosé. 2010. Socially capable conversational tutors can be effective in collaborative learning situations. In *International conference on intelligent tutoring systems*. Springer, 156–164. https://doi.org/10.1007/978-3-642-13388-6_20

[36] Igor Labutov and Hod Lipson. 2012. Humor as circuits in semantic networks. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 150–155. https://doi.org/doi/abs/10.5555/2390665.2390703

[37] Min Kyung Lee, Sara Kielser, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully Mitigating Breakdowns in Robotic Services. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction* (Osaka, Japan) *(HRI '10)*. IEEE Press, 203–210. https://doi.org/10.1109/HRI.2010.5453195

[38] Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, and Yung-Ju Chang. 2020. *A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376209

[39] 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* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 264–275. https://doi.org/10.1145/2901790.2901842

[40] Q. Vera Liao, Muhammed Mas-ud Hussain, Praveen Chandar, Matthew Davis, Yasaman Khazaeni, Marco Patricio Crasso, Dakuo Wang, Michael Muller, N. Sadat Shami, and Werner Geyer. 2018. *All Work and No Play?* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3173577

[41] 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* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[42] Mikael Lundell Vinkler and Peilin Yu. 2020. Conversational Chatbots with Memory-based Question and Answer Generation.

[43] Xiaojuan Ma, Emily Yang, and Pascale Fung. 2019. Exploring Perceived Emotional Intelligence of Personality-Driven Virtual Agents in Handling User Challenges. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 1222–1233. https://doi.org/10.1145/3308558.3313400

[44] Natascha Mariacher, Stephan Schlögl, and Alexander Monz. 2021. Investigating Perceptions of Social Intelligence in Simulated Human-Chatbot Interactions. In *Progresses in Artificial Intelligence and Neural Systems*. Springer, 513–529. https://doi.org/10.1007/978-981-15-5093-5_44

[45] Wen Minlin and Chen Yali. 2017. An Analysis of Self-mockery in Conversational Implicature. (2017). https://doi.org/10.2991/ifmca-16.2017.53

[46] Nicole Mirnig, Susanne Stadler, Gerald Stollnberger, Manuel Giuliani, and Manfred Tscheligi. 2016. Robot humor: How self-irony and Schadenfreude influence people's rating of robot likability. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 166–171. https://doi.org/10.1109/ROMAN.2016.7745106

[47] Nicole Mirnig, Gerald Stollnberger, Manuel Giuliani, and Manfred Tscheligi. 2017. Elements of Humor: How Humans Perceive Verbal and Non-Verbal Aspects of Humorous Robot Behavior. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (Vienna, Austria) *(HRI '17)*. Association for Computing Machinery, New York, NY, USA, 211–212. https://doi.org/10.1145/3029798.3038337

[48] Kellie Morrissey and Jurek Kirakowski. 2013. 'Realness' in chatbots: establishing quantifiable criteria. In *International conference on human-computer interaction*. Springer, 87–96. https://doi.org/10.1007/978-3-642-39330-3_10

[49] Matthijs P Mulder and Antinus Nijholt. 2002. *Humour research: State of the art*. Citeseer.

[50] Anton Nijholt, Andreea I Niculescu, Alessandro Valitutti, and Rafael E Banchs. 2017. Humor in human-computer interaction: a short survey. In *Adjunct conference proceedings interact*. 527–530.

[51] Mohammad Nuruzzaman and Omar Khadeer Hussain. 2018. A Survey on Chatbot Implementation in Customer Service Industry through Deep Neural Networks. In *2018 IEEE 15th International Conference on e-Business Engineering (ICEBE)*. 54–61. https://doi.org/10.1109/ICEBE.2018.00019

[52] Raquel Oliveira, Patrícia Arriaga, Minja Axelsson, and Ana Paiva. 2020. Humor–Robot Interaction: A Scoping Review of the Literature and Future Directions. *International Journal of Social Robotics* (2020), 1–15. https://doi.org/10.1007/s12369-020-00727-9

[53] Zhenhui Peng, Yunhwan Kwon, Jiaan Lu, Ziming Wu, and Xiaojuan Ma. 2019. *Design and Evaluation of Service Robot's Proactivity in Decision-Making Support Process*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi-org.lib.ezproxy.ust.hk/10.1145/3290605.3300328

[54] Zhenhui Peng and Xiaojuan Ma. 2019. A survey on construction and enhancement methods in service chatbots design. *CCF Transactions on Pervasive Computing and Interaction* 1, 3 (2019), 204–223. https://doi.org/10.1007/s42486-019-00012-3

[55] Zhenhui Peng, Xiaojuan Ma, Diyi Yang, Ka Wing Tsang, and Qingyu Guo. 2021. Effects of Support-Seekers' Community Knowledge on Their Expressed Satisfaction with the Received Comments in Mental Health Communities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3411764.3445446

[56] Dulce G Pereira, Anabela Afonso, and Fátima Melo Medeiros. 2015. Overview of Friedman's test and post-hoc analysis. *Communications in Statistics-Simulation and Computation* 44, 10 (2015), 2636–2653. https://doi.org/10.1080/03610918.2014.931971

[57] Maria João Pereira, Luísa Coheur, Pedro Fialho, and Ricardo Ribeiro. 2016. Chatbots' Greetings to Human-Computer Communication. arXiv:1609.06479 [cs.HC]

[58] Silvia Quarteroni and Suresh Manandhar. 2009. Designing an interactive open-domain question answering system. *Natural Language Engineering* 15, 1 (2009), 73–95. https://doi.org/10.1017/S1351324908004919

[59] Byron Reeves and Clifford Nass. 1996. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Pla. *Bibliovault OAI Repository, the University of Chicago Press* (01 1996).

[60] Peter Salovey and John D Mayer. 1990. Emotional intelligence. *Imagination, cognition and personality* 9, 3 (1990), 185–211.

[61] Ari Schlesinger, Kenton P. O'Hara, and Alex S. Taylor. 2018. *Let's Talk About Race: Identity, Chatbots, and AI.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173889

[62] Joseph Seering, Michal Luria, Connie Ye, Geoff Kaufman, and Jessica Hammer. 2020. *It Takes a Village: Integrating an Adaptive Chatbot into an Online Gaming Community.* Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376708

[63] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (Phoenix, Arizona) *(AAAI'16)*. AAAI Press, 3776–3783. https://doi.org/10.48550/arXiv.1507.04808

[64] Annika Silvervarg and Arne Jönsson. 2013. Iterative development and evaluation of a social conversational agent. In *6th International Joint Conference on Natural Language Processing (IJCNLP 2013), 14-18 October 2013, Nagoya, Japan*. 1223–1229.

[65] Yiping Song, Rui Yan, Cheng-Te Li, Jian-Yun Nie, Ming Zhang, and Dongyan Zhao. 2018. An Ensemble of Retrieval-Based and Generation-Based Human-Computer Conversation Systems. (2018). https://doi.org/10.24963/ijcai.2018/609

[66] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*. https://doi.org/10.48550/arXiv.1612.03975

[67] Moonyoung Tae and Joonhwan Lee. 2020. The Effect of Robot's Ice-Breaking Humor on Likeability and Future Contact Intentions. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 462–464. https://doi.org/10.1145/3371382.3378267

[68] Carlos Toxtli, Andrés Monroy-Hernández, and Justin Cranshaw. 2018. Understanding chatbot-mediated task management. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–6. https://doi.org/10.1145/3173574.3173632

[69] Sheldon Ungar. 1984. Self-mockery: An alternative form of self-presentation. *Symbolic Interaction* 7, 1 (1984), 121–133. https://doi.org/10.1525/si.1984.7.1.121

[70] Sarah Theres Völkel, Daniel Buschek, Malin Eiband, Benjamin R. Cowan, and Heinrich Hussmann. 2021. Eliciting and Analysing Users' Envisioned Dialogues with Perfect Voice Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 254, 15 pages. https://doi.org/10.1145/3411764.3445536

[71] Peter Wallis and Emma Norling. 2005. The Trouble with Chatbots: social skills in a social world. *Virtual Social Agents* 29 (2005), 29–36.

[72] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From Human-Human Collaboration to Human-AI Collaboration: Designing AI Systems That Can Work Together with People. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3334480.3381069

[73] Julia Wilkins and Amy Janel Eisenbraun. 2009. Humor theories and the physiological benefits of laughter. *Holistic nursing practice* 23, 6 (2009), 349–354. https://doi.org/10.1097/hnp.0b013e3181bf37ad

[74] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).

[75] Robert F Woolson. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* (2007), 1–3.

[76] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. *A New Chatbot for Customer Service on Social Media.* Association for Computing Machinery, New York, NY, USA, 3506–3510. https://doi.org/10.1145/3025453.3025496

[77] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 55–64. https://doi.org/10.1145/2911451.2911542

[78] Xi Yang and Marco Aurisicchio. 2021. Designing Conversational Agents: A Self-Determination Theory Approach. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 256, 16 pages. https://doi.org/10.1145/3411764.3445445

[79] Nancy A Yovetich, J Alexander Dale, and Mary A Hudak. 1990. Benefits of humor in reduction of threat-induced anxiety. *Psychological Reports* 66, 1 (1990), 51–58. https://doi.org/10.2466/pr0.1990.66.1.51

[80] Lina Zhou, Liwei Dai, and Dongsong Zhang. 2007. Online shopping acceptance model-A critical survey of consumer factors in online shopping. *Journal of Electronic commerce research* 8, 1 (2007), 41.

[81] Darius Zumstein and Sophie Hundertmark. 2017. CHATBOTS–AN INTERACTIVE TECHNOLOGY FOR PERSONALIZED COMMUNICATION, TRANSACTIONS AND SERVICES. *IADIS International Journal on WWW/Internet* 15, 1 (2017).