Delft University of Technology

Transparent Conversational Agents
The Impact of Capability Communication on User Behavior and Mental Model Alignment

Reimann, Merle M.; Kunneman, Florian A.; Oertel, Catharine; Hindriks, Koen V.

DOI
10.1145/3719160.3736629

Licence
CC BY

Publication date
2025

Document Version
Final published version

Published in
CUI 2025 - Proceedings of the 2025 ACM Conference on Conversational User Interfaces

Citation (APA)
Reimann, M. M., Kunneman, F. A., Oertel, C., & Hindriks, K. V. (2025). Transparent Conversational Agents: The Impact of Capability Communication on User Behavior and Mental Model Alignment. In J. Sin, E. Law, J. Wallace, C. Munteanu, & D. Korre (Eds.), *CUI 2025 - Proceedings of the 2025 ACM Conference on Conversational User Interfaces* (pp. 1-12). Article 48 Association for Computing Machinery (ACM). https://doi.org/10.1145/3719160.3736629

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Transparent Conversational Agents: The Impact of Capability Communication on User Behavior and Mental Model Alignment

Merle M. Reimann
Vrije Universiteit Amsterdam
Amsterdam, Netherlands
m.m.reimann@vu.nl

Florian A. Kunneman
Utrecht University
Utrecht, Netherlands
f.a.kunneman@vu.nl

Catharine Oertel
TU Delft
Interactive Systems/Interactive Intelligence
Delft, Netherlands
c.r.m.m.oertel@tudelft.nl

Koen V. Hindriks
Vrije Universiteit Amsterdam
Amsterdam, Netherlands
k.v.hindriks@vu.nl

Abstract

When a user interacts with a conversational agent for the first time, they may not be aware of the agent's capabilities, leading to suboptimal use or interaction breakdowns. To avoid a mismatch with the actual capabilities, the agent's capabilities have to be made transparent to the user. To investigate whether communication of an agent's capabilities during interactions enhances transparency and improves the user's mental model, we conducted a user study with 56 participants. Each participant had three speech-based interactions with an agent that communicated its capabilities or an agent that did not. Our results suggest that the communication led to a change in user behavior with significantly longer utterances. However, the users' mental models of the agent's capabilities were not significantly different between the conditions. Participants were able to significantly improve their knowledge of the agent's capabilities by aligning their mental model over time in both conditions.

CCS Concepts

• **Human-centered computing** → *Empirical studies in HCI*; **Natural language interfaces**; *Collaborative interaction*; *Empirical studies in interaction design.*

Keywords

Conversational Agents, Dialogue Management, Mental Models, Capability Communication, Transparency

ACM Reference Format:
Merle M. Reimann, Florian A. Kunneman, Catharine Oertel, and Koen V. Hindriks. 2025. Transparent Conversational Agents: The Impact of Capability Communication on User Behavior and Mental Model Alignment. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25), July 08–10, 2025, Waterloo, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3719160.3736629

This work is licensed under a Creative Commons Attribution 4.0 International License.
CUI '25, Waterloo, ON, Canada
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1527-3/25/07
https://doi.org/10.1145/3719160.3736629

1 Introduction

While interactions with conversational agents have become more common, not all agents have the same communication capabilities. For example, some agents might understand references to earlier parts of the conversation, while others may not. For a user, this means that it is inherently unclear what the agent can actually understand and do if no additional information is provided. This can lead to a mismatch between the actual capabilities of the agent and the user's assumptions about them. These user assumptions can be seen in the user's mental model which is a conceptual framework of entities that people form and change based on expectations and experiences [9].

The mismatches between the user's conception of an agent's capabilities and the actual capabilities can occur either because expected capabilities are not there or because the user does not know about capabilities the agent has. The first type of mismatch will typically lead to a misunderstanding and could ultimately lead to a breakdown. For the second type of mismatch, the efficiency or quality of the interaction is not at the level that it could be if the user was aware of all the capabilities. Although communicating the agent's capabilities to the user could make the interaction more robust and efficient, there is no single standard way of doing this.

For robots, additional information that can inform the user's expectations about the robot's capabilities can come in the form of the robot's embodiment and the context in which it is placed [5, 14]. However, the expectations created by the embodiment of the robots and their behavior do not necessarily align with the actual capabilities that the robot has [17]. For chatbots, there is no embodiment that could be used to influence the user's expectations of the agent's capabilities. Instead, past experience and context, as well as observations made during the interaction, shape the mental model of the user.

Valério et al. [23] identify different methods of how the visual user interface can be used to convey the capabilities of a conversational agent to a user. A popular method is to limit the user's input methods by providing fixed input options that limit the navigational freedom [23]. In an interaction in an open domain scenario, where the interaction is not limited to a specific context but could be about various topics, such a restriction is more difficult. Especially for agents that are solely relying on speech to communicate, the communication of capabilities has to be integrated into the conversation itself. For such systems, the user input cannot be easily limited, and the agent has to communicate the capabilities along its task-related output. Task-related outputs are utterances that are directly relevant for task completion (e.g. giving the directions in a navigational task). In contrast to that, the general capabilities (e.g. repeating utterances in case of misunderstanding) are not necessarily linked to the concrete task. The capability that is communicated and the moment at which this is done should hence be chosen carefully to not negatively impact the task performance. How conversational agents can communicate their capabilities to a user during a speech-based interaction and the effects of this capability communication on the user's behavior and perceived conversation quality have not been explored in-depth, according to our knowledge.

With our work we expand the understanding of user expectations in regard to conversational capabilities of conversational agents and how those expectations change over time. Additionally, we present repair mechanisms that facilitate the capability communication of the agent in order to reduce mismatch between the agent's capabilities and the user's mental model. To analyze this capability communication, we conducted a user study with 56 participants, who had three interactions each and recorded their capability knowledge after each interaction. With our work, we aim at answering the following research questions:

1. How does a conversational agent's communication about its capabilities influence
   (a) the perceived conversation quality?
   (b) the user's mental model of the agent's capabilities?
   (c) the user's usage of the agent's capabilities?
2. How does learning over time affect the user's mental model of the agent's capabilities?
3. How do a user's mental model of the agent's capabilities and perceived conversation quality relate?

2 Background

To understand how capability communication can be used to improve the interaction with the agent, we will discuss some concepts relevant to our work in this section. First, we introduce the concept of mental models (see Section 2.1), since we are aiming to adjust the user's mental model with our capability communication. The adjustment has two goals: 1) reducing the user's uncertainty regarding the agent's capabilities and 2) correcting a possible misalignment with the actual capabilities through transparency. Methods that have been used to move towards this goal will be discussed in the following sections, by introducing the Uncertainty Reduction Theory (see Section 2.2) and methods of capability communication (see Section 2.3).

2.1 Mental Models

Humans form mental models of their environment and entities they are interacting with, including conversational agents. Those models are used to make predictions and explain and understand observed phenomena [9]. In the context of conversational agents, mental models shape expectations about the agent's behavior and capabilities. Mental models are influenced by explanations and can change over time through experience [13]. In cases where a person has no prior experience, they can transfer their knowledge they have about a similar concept or entity to the unknown one as an initial mental model [8]. This means that even users who use a system for the first time can have certain – not necessarily correct – expectations regarding the agent's behavior and capabilities.

Violations of a mental model can negatively impact the user's evaluation of the system [8]. However, Kulesza et al. [12] found that users who changed their mental models the most during the interaction to fit with the actual capabilities were more likely to be satisfied with the interaction.

Mental models are also relevant for explainable AI, as they offer an option to assess the knowledge and assumptions of users about the system they are interacting with [22]. We can also use the users' mental models to evaluate how much they know about the agent's capabilities.

2.2 Uncertainty Reduction Theory

Due to the high number of available conversational agents, it is common that a person is interacting with a specific agent for the first time. When people first interact without knowing anything about each other, a central aspect of this initial interaction is to learn about each other to reduce uncertainty about the interaction and increase the predictability of the behavior [1]. Uncertainty Reduction Theory further suggests that the uncertainty decreases through further interaction and that high levels of uncertainty increase the amount of information seeking in interactions.

Additionally, Berger and Calabrese [1] state "[s]imilarities between persons reduce uncertainty, while dissimilarities produce increases in uncertainty". In human-agent interaction the human is not just interacting with a person that is different from them, but with a completely different entity that can even have unknown features or abilities. Liu [16] observed a similar effect for human-agent interaction, suggesting that the concept of the uncertainty reduction theory might generalize to human-agent interaction.

This problem of uncertainty is also relevant where a lack of knowledge about the system can lead to uncertainty regarding the robot's or agent's capabilities [8, 17]. One possible strategy to overcome this discrepancy between the assumed and actual capabilities is to communicate the agent's capabilities to the user to help align their mental model with the actual capabilities, which is what we aim at with our capability communication strategies.

2.3 Capability Communication

Machines that interact in a human-like manner can evoke the feeling that they do in fact have human like capabilities, even though they do not [19]. The discrepancy between actual and assumed capabilities of social robots and the concomitant problems are discussed in [3] and [17], who also point out that there is no optimal solution for the discrepancy minimization yet. Valério [23] analyze three common chatbots regarding their capability communication and identify eleven strategies used. However, the commonly used strategies mostly involve multimodal interfaces, such as visual menus supporting the user. In contrast to those strategies, we focus on how to integrate the capability communication directly into the spoken interaction in the form of repair mechanisms.

Since humans vary in their interaction styles and needs, transparent communication about the capabilities is relevant for increasing the inclusiveness as well [20]. Kirschthaler et al. [11] show that increasing the discoverability of capabilities in conversational agents improves the user experience. Especially contextualized help, which highlights capabilities when they are actually relevant, improves discoverability and learnability [4]. Improving the transparency of systems, for example, by contextualizing open-ended prompts and making possible actions more explicit, helps users navigate the interaction with the system [10]. Another option of navigating the interaction with the system is by having the system guide it through questions that directly address the next state of the interaction [6].

Transparency also plays a role in human-robot interaction, where it is evaluated by looking at the robustness of the interaction, the trust in the agent and the interaction efficiency [24]. Robustness of an interaction evaluates how easily an interaction fails if something goes wrong. The degree of soundness and completeness of explanations influences the creation and maintenance of mental models, with completeness being more important [13].

The problems caused by lack of transparency and capability communication can be summarized into two categories: 1) the user overestimates the capabilities of the system and 2) the user misses capabilities the system would be able to perform [25]. Based on those two categories, which are not mutually exclusive, we designed repair strategies that allow us to communicate the actual capabilities of the agent in a targeted manner.

3 Hypotheses

To learn more about how the agent's communication about its capabilities is perceived by the user and influences their behavior, we target the following research questions in our paper:

(1) How does a conversational agent's communication about its capabilities influence
  (a) the perceived conversation quality?
  (b) the user's mental model of the agent's capabilities?
  (c) the user's usage of the agent's capabilities?

Looking at the influence of the repair communication shows us whether the employed strategy is not only suitable for communicating the capabilities, but also its effects on the quality of the interaction. This is important since an interaction that is of low quality could lead to the user abandoning the interaction early or avoiding engaging in another interaction with the agent in the future. The perceived conversation quality and user model of the agent's capabilities come from questionnaire data, whereas the usage of the capabilities is based on observational data. The repair communication about the agent's capabilities provides the user with more information and explanations about the failure and ways to prevent it than a generic response would. Therefore, we hypothesize that repair communication has the following effect:

H1: The agent's capability communication improves:
a) perceived conversation quality
b) the user's mental model of the agent's capabilities
c) the user's usage of the agent's capabilities.

These hypotheses are based on the observation by Kirschthaler that increased discoverability of capabilities in conversational agents improves the user experience [11]. Since capability communication also aims at increasing the discoverability, we assume that this is also the case for our strategies.

(2) How does learning over time affect the user's mental model of the agent's capabilities?

We let the participants interact with the agent multiple times, to see how the user's mental model of the agent's capabilities develops over time. During the interaction, the user can experience and explore the capabilities of the agent, regardless of the repair strategy. However, if the communication repair strategy is used, the agent can facilitate this process and also encourage the user to try capabilities they were not aware of. However, for both conditions, we hypothesize that:

H2: A learning effect can be observed regarding the user's mental model of the agent's capabilities.

We expect that users are able to learn the agent's capabilities, leading to a better alignment with the actual capabilities. This hypothesis is in line with the uncertainty reduction theory [1], since the repeated interactions allow participants to infer information about the agent's behavior and capabilities by observing and testing it multiple times.

(3) How do a user's mental model of the agent's capabilities and perceived conversation quality relate?

The interaction of the possible learning effect and perceived conversation quality can provide us insights into how the uncertainty reduction over time can impact the user's perception of the agent. While an improved model enables the user to use the capabilities of the agent better, it might also be frustrating for the user if the assumptions have to be updated over and over, especially if the capabilities are worse than expected, and hence we hypothesize that:

H3: There is a positive relation between the alignment of the user's model with the agent's capabilities and the perceived conversation quality.

This hypothesis is motivated by the findings of Kulesza et al. [12], who observed higher interaction satisfaction for people who were able to adjust their mental model to the actual capabilities.

4 Methods

To investigate the presented research questions and related hypotheses, we ran an experiment, which we will describe in the following sections.

4.1 Participants

We recruited a total of 56 participants using flyers and word-of-mouth advertisements. Due to technical difficulties during the interaction, four participants had to be excluded, leading to a total of 52 participants being included in the data (age: 18-33 (M=24.23, SD=3.88); 19 female, 28 male, 4 diverse, 1 prefer not to say). The number of participants was determined using an a-priori power analysis ($\alpha = 0.05$, $\beta = 0.8$, $d = 0.4$) using G*Power [7], which led to 52 participants needed. All participants were aged 18 or older,

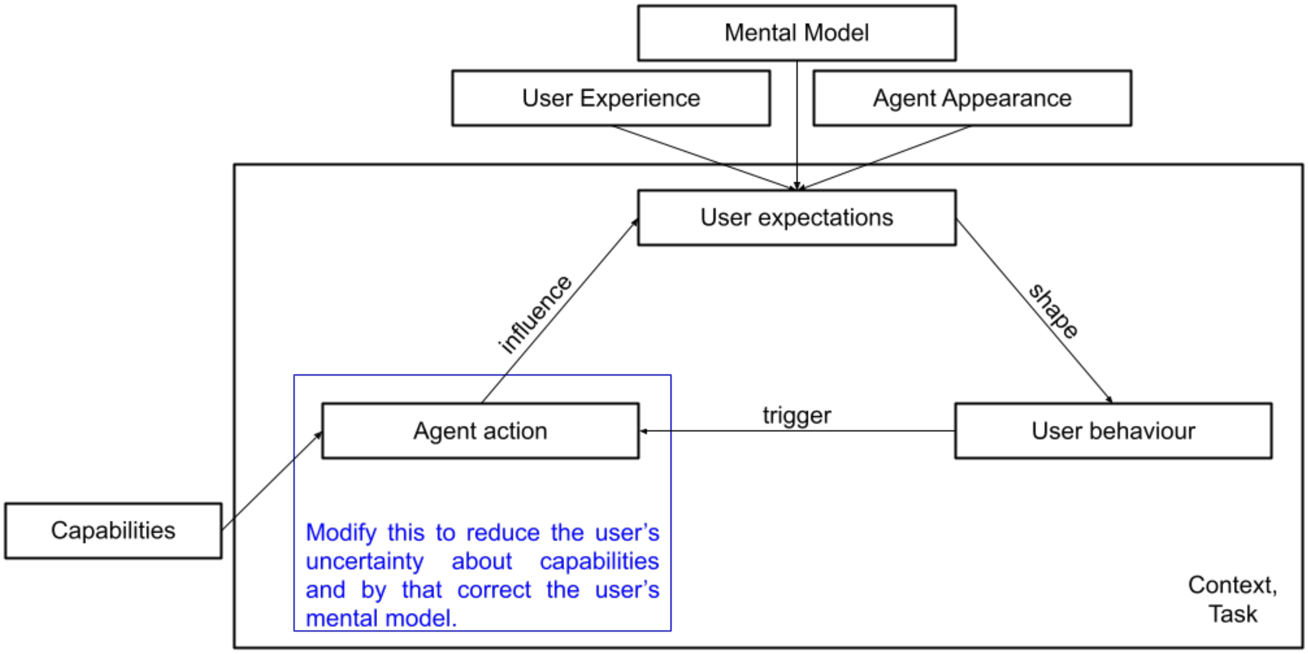

Figure 1: With our repair strategies, we modify the agent's actions in order to influence the user's mental model of the agent and their expectations and behavior.

spoke English fluently and had no severe speech or hearing impairments. Participants received a 10€ gift card as a thank you for their participation. Before the experiment, all participants were asked to rate their prior experience with speech-based conversational agents (None at all (14 participants), A little (15), A moderate amount (17), A lot (4), A great deal (2)).

4.2 Agent

The agent used is a recipe recommendation agent that filters a database of recipes based on user preferences and can display fitting recipes to the user. The user can select recipes for closer inspection and decide based on that whether they want to select this recipe or continue looking for others. We decided to use this agent, since recipe recommendation is a semi-structured task, with a clear conversation flow that still allows for some deviations, to allow for exploration of the agent's capabilities. For this the vast knowledge space of the agent for recipe recommendations is also an advantage, since it allows for a variety in capabilities that could be used.

We used Google speech recognition and Dialogflow for recognizing what the user was saying and classifying it into a set of intents. The agent uses model-based dialogue management with conversational patterns, following the Intent-Entity-Context-Response (IECR) framework [18]. Conversational patterns are based on common conversation structures and adjacency pairs (e.g., question-answer, greeting-greeting). Each intent extracted from the user utterances is matched with the currently active conversation patterns and those patterns the user could initiate to determine the next dialogue move of the agent. If not only an intent, but also entities are returned, the agent uses those entities to perform the desired action, for example, filter based on given ingredient preferences. To make the interaction sound less static, we vary the agent utterance by forwarding the response utterance the agent selects, together with the conversation history, to GPT-4o-mini, which is prompted to return a fitting reformulation of the utterance.

The baseline agent follows a simple repair strategy. In cases of misunderstandings, the agent tells the user that it did not understand the utterance, without going into detail what it understood, and asks the user to repeat their last utterance.

Capability communication agent. The capability communication agent uses repair to reduce the uncertainty in the interaction (see Figure 1). We decided to integrate the capability communication into repair so that we can target the two problems pointed out by Yankelovich [25]: Overestimation and missed capabilities. Integrating capability communication into repair mechanisms increases the transparency of the agent by providing information on what the agent's actual capabilities are only when that information is actually needed.

In case the agent understands an intent but is unable to match it to an existing pattern, it will inform the user which intent it understood and what would have been the intent, which would correspond to the happy flow of the conversation. The happy flow describes the optimal flow of the conversation for the given task and context. If the agent is unable to recognize the intent at all, it generates a repair utterance, referencing the misunderstood user utterance, asking the user to rephrase and resolve the misunderstanding. The agent is also able to detect inefficiencies in the conversation, for example, when the user is adding ingredients one by one, instead of adding multiple at once. If the user has not previously demonstrated knowledge of this feature, by adding multiple ingredients at once, the agent will inform the user about the feature after three ingredients have been added separately. To avoid annoying the user through repetitions, the agent will not mention it again afterwards.

The capability communication agent is also using GPT-4o-mini for reformulating the utterances, with an emphasis on using the conversation history, to embed the repair in a helpful manner. All prompts used for the agent can be found in the Appendix (see Table 7).

4.3 Design

The experiment has two conditions and a between groups design. The two conditions differ in the agent's communication strategies regarding its capabilities. The first agent ("basic agent") was used as a baseline and had basic repair strategies, without a focus on capability communication. The second agent ("capability communication agent") used repair strategies to communicate its conversational capabilities. The independent variable in our experiment was the type of agent used, while the dependent variables are perceived conversation quality, quantitative behavioral measures (e.g. number of turns), and the mental model alignment (capabilities known by the user).

Each user was asked to have three interactions with the agent. The perceived conversation quality was measured using a questionnaire after the first and third interaction, while the quantitative measures were extracted from interaction logs recorded during the interaction. To assess the users' capability knowledge, they were asked to fill in a questionnaire after each interaction block. The perceived conversation quality questionnaire was only assessed after the first and the last interaction, since the focus was on the capability knowledge, and we did not want to overload the participants with questionnaires, to avoid them losing concentration.

4.4 Data collection

The perceived conversation quality was measured using subscale 3 of the BOT Usability Questionnaire (4 items, $\alpha = 0.86$) [2].

For the capabilities used, we created a list of existent and nonexistent capabilities (see table 1). With this list of 22 capabilities, we then ran an online prestudy using Prolific (www.prolific.com). A total of 50 participants were asked to rate for each capability how likely they think it is that the agent has the capability. Before rating the capabilities, they were shown a picture of the interface and listened to a small audio fragment of the agent's voice. Based on the results of the prestudy, we selected six capabilities that users were asked to rate during the actual study. We decided to reduce the number to 6, since it allowed us to have a combination of existent and non-existent capabilities with different ratings, while not fatiguing the participant with too many questions. The capabilities were chosen in such a way that two of the existing capabilities had high ratings (3.32/4, 3.36/4), while two had lower ratings (2.57/4, 2.94/4). For the nonexistent capabilities, one with a low (2.64/4) and one with a high (3.42/4) rating were selected. We decided to use this approach for creating the capabilities list to also take the initial expectations of users into account.

The questionnaire answers for the capability rating were on a 4-point Likert scale with an additional "I do not know". The perceived conversation quality was a 5-point Likert scale. The interaction logs used contain the transcripts of user and agent utterances, the conversation history and agent memory, as well as the intents extracted from the user utterances. For our analysis, we focused on the user utterances and their intents.

4.5 Procedure

The experiment took place in person. After reading and accepting the consent form, the participants were asked to interact with the agent and find a recipe. They were told that the agent is a speech-based agent that can help them to find a recipe and that their task is to find a recipe they are satisfied with. Additionally they were asked to speak as clearly as possible to minimize speech recognition problems. After completing one interaction, the participants were asked to fill in the perceived conversation quality questionnaire and rate how well they think the agent can perform the 6 capabilities selected after the online prestudy. After the second interaction, the participants were asked to rate the capabilities again. Following the final interaction, they filled in the perceived conversation quality again, and provided the capability ratings, this time in combination with explanations on why they decided to rate like this.

4.6 Analysis

For determining whether differences between conditions were significant, we first checked whether all conditions for an ANOVA were met and used a Mann-Whitney U test, if it was not the case.

To compare different time points during an interaction, we checked whether the conditions for an one-way repeated measures ANOVA were met and used a Friedman test as a non-parametric alternative, if they were not.

For post-hoc corrections, we used the Benjamini-Hochberg correction.

4.6.1 *Capability Knowledge.*

For the capability knowledge, we differentiate between the (mis-)alignment of the mental model with the actual capabilities and the level of uncertainty of the user. For calculating the mental model alignment we first considered coding the "I do not know" answers as being maximally far from the correct mental model, but then decided to not include those answers in the alignment, since it would mix the uncertainty with the alignment. Therefore, we decided to include only the ratings and grouped them into a binary classification by clustering the "I think" and "I know" answers together. To be able to calculate the distance of the user's mental model of the capabilities and the actual ones, we coded the answers with values from 1 to 4. To calculate the alignment of the user's mental model, we used

Table 1: The capabilities the participants were asked to rate based on how well the agent presumably could perform the capability. The capability descriptions marked in bold were selected for the experiment.

Exists	Does not exist
Filter on dietary preference	**Filter based on season**
Filter on inclusion of ingredients	Filter based on recipe steps
Filter on cuisine	**Read recipe step by step**
Filter on recipe complexity	Suggest replacements for ingredients
Filter on time needed	Suggest a wine to go with the food
Filter on exclusion of ingredients	Repeat/rephrase the last utterance
Suggest a random recipe	Refer to previous conversations
Display a list of recipes matching all constraints thus far	
Remove all preferences to filter recipes on	
Display a specific recipe	
Remove a single preference used to filter recipes on	
Explain what the agent can do	
Ask the user for additional preference information	
Ask the user for clarification/repetition	
Remove all preferences to filter recipes on	

$$\text{Alignment score} = \sum_{i=1}^{n} \begin{cases} 1, & \text{if } R_i == 3 \text{ or } 4 \text{ and } i \text{ in } C, \\ & \text{or } R_i == 1 \text{ or } 2 \text{ and } i \text{ not in } C \\ -1, & \text{if } R_i == 1 \text{ or } 2 \text{ and } i \text{ in } C, \\ & \text{or } R_i == 3 \text{ or } 4 \text{ and } i \text{ not in } C \\ 0 & \text{if ``I do not know'' was chosen} \end{cases} \tag{1}$$

where R_i is the transformed (on a 1-4 scale, with 1 as "I know that the agent cannot do this" and 4 being "I know the agent can do this") rating given by the participant for capability i, C are all agent capabilities and $n(= 6)$ is the number of capabilities included in the questionnaire. With our list of six capabilities per interaction, the maximum error is -6 and a perfectly aligned model corresponds to a score of 6.

5 Results

We analyzed both subjective user ratings in form of questionnaire answers, and objective metrics extracted from the logs generated during the interaction.

5.1 Interaction Example

The interaction started when the user pressed the start button of the agent. The agent then greeted the user and asked them what kind of recipe they were looking for. This started the first phase of the interaction which was the selection of the recipe. The user now had the chance to either name a recipe, state their requirements (e.g. dietary restrictions, ingredient in- and exclusion, complexity) or ask for a recommendation. Based on the user's request the agent then filtered its database and automatically displayed suggestions if maximally 20 recipes were left. If the number of fitting recipes was smaller than 100, the user could also request to see them. The agent then asked the user whether they want to add or remove any requirements to refine the search. Once the user was interested in a displayed recipe, they could enter the recipe confirmation stage. They did this, by asking the agent to show the recipe to them and then were able to decide whether they want to cook it, which ended the interaction, or if they want to return to the recipe selection stage.

The following transcript shows an interaction (P24, Interaction 1):

> **Agent:** What do you enjoy most in a recipe?
> **User:** I like pasta
> **Agent:** All the remaining recipes include pasta. Would you like to add another preference?
> **User:** hey Ya
> **Agent:** I didn't quite catch that. Could you clarify?
> **User:** yes I am also a vegetarian
> **Agent:** All the remaining recipes follow a vegetarian diet. Would you like to add another preference?
> **User:** I like cheese
> **Agent:** Got it! What else would you like to add or remove? You can specify multiple preferences at once.
> **User:** I prefer tomato based sauces with basil
> **Agent:** It looks like none of my recipes match your preferences. Could you please consider removing one?
> **User:** yes remove with base sauce
> **Agent:** Here are some recipes that match your preferences.
> **User:** Bens summer tomato pasta
> **Agent:** Ben's Summer Tomato Pasta is a fantastic choice! Would you like to check the recipe and let me know if you want to cook it?
> **User:** looks good I want to cook it
> **Agent:** Did you manage to complete the task successfully? Hope you have a great day! Goodbye!

In this interaction the user was satisfied with the first recipe they had a closer look at. After the user added two ingredients one-by-one the agent informed the user that they could also add multiple ingredients at once, which the user then did. In this specific case it led to a combination of the requirements that did not match any recipe in the database leading to the user being requested to adjust their search. The length of the interaction depended on the user's requirements, interaction style, and also on speech recognition problems, which could lead to additional repairs being needed.

5.2 Questionnaire

5.2.1 Perceived Conversation Quality. Since the data for the BUS questionnaire after the first interaction were not normally distributed, we used the nonparametric Mann-Whitney U test. The difference between the conditions was not significant ($p = .56$).

The questionnaire data after all three interactions were normally distributed and homoscedasticity was given, which is why we used the parametric one-way ANOVA, which also showed no significant difference between the conditions ($p = .84$).

Capability knowledge between conditions. Calculating the average mental model alignment scored for the two conditions over the three interactions showed no significant difference between the conditions after a post-hoc correction for either of the interactions ($p = .70$, $p = .51$, $p = .70$). While the mean scores for the first two interactions were slightly higher in the baseline condition, this changed for the last interaction (see Table 2).

Table 2: The mean and standard deviation of the mental model alignment scores of the two conditions for the three interactions.

Interaction	Baseline	Capability Communication
1	2.65 (1.87)	2.38 (1.72)
2	3.00 (1.94)	2.46 (1.68)
3	3.69 (2.54)	3.73 (1.87)

For the rating of the capabilities, the scale included an "I do not know" option, which was used for 37 (baseline) and 26 (capability communication) items after the first interaction. The number decreased for both conditions for the second and third interaction (see Table 3). Calculating the mean percentage of "I do not know" answers per participant, also showed higher uncertainty for participants in the baseline condition (see Figure 2), however, the differences were not significant ($p = .45$).

Table 3: The number of times "I do not know" was given as an answer.

	Interaction		
Condition	First	Second	Third
Baseline	37	28	16
Capability Communication	26	20	13

Capability knowledge over time. Since the condition of following a normal distribution was not given, we performed a Friedman test instead of a repeated measures ANOVA. The Friedman test showed significant changes in the mental model's alignment over time for the capability communication condition ($p < .05$), but also for the baseline condition ($p < .05$). The decrease in the difference between the mental model and the actual capabilities per participant (based on equation 1) is visualized in Figure 3. The average mental model alignment error was reduced from $M = 7.31$ ($SD = 2.29$) to $M = 6.54$ ($SD = 2.35$) and then to $M = 4.96$ ($SD = 2.58$) in the capability communication condition and from $M = 7.85$ ($SD = 2.87$) to $M = 6.92$ ($SD = 2.97$) and then to $M = 5.54$ ($SD = 3.05$) in the baseline condition.

The capability knowledge also depends on whether a participant has already encountered the capability (see Figure 4). For capabilities that the agent did not have, "I do not know" was the most common answer (Total: 35.58%, Baseline: 41.67%, Capability Communication: 29.49%), whereas it was "I think the agent can do this" (Total: 40.74%, Baseline: 35.53%, Capability Communication: 45.35%) for the capabilities that were there, but had not yet been encountered by the participant, and "I know the agent can do this" (Total: 70.13%, Baseline: 66.10%, Capability Communication: 74.34%) for capabilities that were encountered.

5.2.2 Capability Knowledge and Perceived Conversation Quality. To analyze whether there is a relationship between the alignment error of the mental model and perceived conversation quality, we performed a linear regression. While the result was significant ($p < .05$) the R-squared showed that only a small part of the variance can be explained by the model ($r2 = 0.09$). The same could be observed with the number of misunderstandings in relation to the perceived interaction quality ($p < .05$, $r2 = 0.05$).

5.3 Qualitative Analysis

5.3.1 Comments on reasoning behind scores. After the third interaction, participants were asked to provide additional information on why they think the agent has / does not have a capability, while providing the ratings. To find common reasons behind believing a capability to be present or absent, all reasons provided were labeled as one of five categories (see Table 4). The labeling was done by all authors and each answer was labeled independently by at least two people to ensure consistency. We had a total of three different rater combinations, and Cohen's Kappa [15] showed substantial agreement ($\kappa = 0.756$), moderate agreement ($\kappa = 0.566$) and fair agreement ($\kappa = 0.223$) after the first round. We then met and discussed the disagreements and tried to reach an agreement wherever possible.

After the last interaction, participants were not only asked to rate whether they think the agent has a capability, but also to give a reason for their rating. We analyzed those reasons by categorizing them into 5 categories (see Table 4). By discussing initial disagreements in our annotation we were able to classify 294 reasons given into one of the categories (see Table 5), while 18 reasons were classified as belonging to two categories.

Reasoning about the agent's capabilities based on other capabilities was often referring to capabilities that were expected to work in a similar manner.

> "The agent was able to find recipes based on the time of having the meal on the day. So, it should also be possible enough to filter out recipes based on seasons."
> - P11

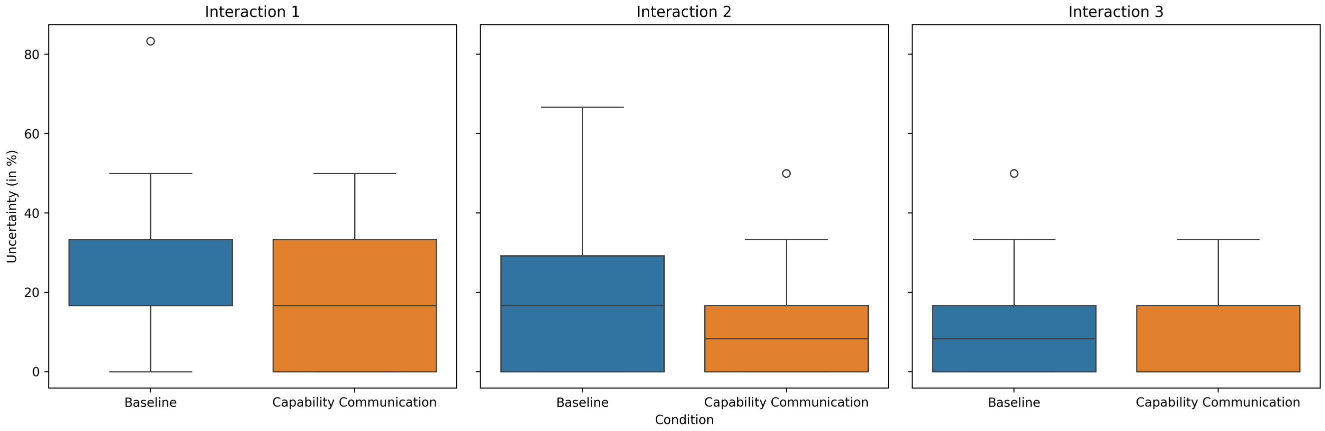

Figure 2: The percentage of "I do not know" answers per condition.

Table 4: The categories for the annotation of the comments providing reasoning for their rating of the agent's capabilities.

Category	Description
Encountered during interaction	The participant mentions that they noticed that capability, or they didn't notice it. The comments did not refer to explicitly testing for a capability nor were any additional explanations of the capability provided.
Explicitly tested	The participant mentions that they explicitly tested for this capability.
Inferred from other capabilities	The participant infers presence of a capability from other capabilities.
Explanation of the capability / its limitations	The participant is explaining the capability they noticed or provides more information about the (limitations of the) capability.
Other	Everything else not fitting any other categories. Includes empty answers.

Table 5: How many comments were classified into each category.

Category	#comments
Encountered during interaction	148
Explicitly tested	29
Inferred from other capabilities	16
Explanation of the capability / its limitations	50
Other	51

51 comments fell into the "Other" category, with most of them being that they do not know, since they did not ask anything related to the capability.

> "I didn't try to filter the ingredients based on the season so I don't know if it is capable of doing that." - P2

It also contained people not answering the question. Some participants also reasoned based on the task difficulty or that the question was included in the questionnaire.

> "I assume that it can also filter based on season, since youre [sic] asking this question." - P10

5.4 Log Data

To gain insights into the user behavior, we analyzed the log data that the agent created during the interaction. The logs included information about the session history, including time, the intents used, the transcripts of the user utterances and the agent's responses.

Based on that information, we were then able to compute the number of intents used and how many intents different from each other were used, as well as sentence length and the number of repairs by the agent initiated after a misunderstanding or mismatch in the interaction (see Table 6).

Although the baseline had on average more intents than the Capability Communication condition, the difference was not significant ($p = .27$). However, a Mann-Whitney U test showed significant differences ($p < .0.05$) in sentence length after a Benjamini-Hochberg correction, with longer sentences in the capability communication condition.

Table 6: Different metrics extracted from the log data. The *Total Intents* include intents that occur multiple times, whereas the *Different Intents* only count each intent type once per interaction. *Sentence Length* is measured in words and *Misunderstanding* and *Mismatch* are two types of problems which trigger repair strategies the agent has.

Condition	Total Intents	Different Intents	Sentence Length	Misunderstanding	Mismatch
Baseline	17.29 (12.42)	6.19 (1.63)	4.80 (1.81)	3.94 (4.43)	1.19 (1.72)
Capability Communication	13.92 (6.52)	6.23 (1.24)	5.61 (2.26)	2.62 (2.69)	0.83 (1.22)

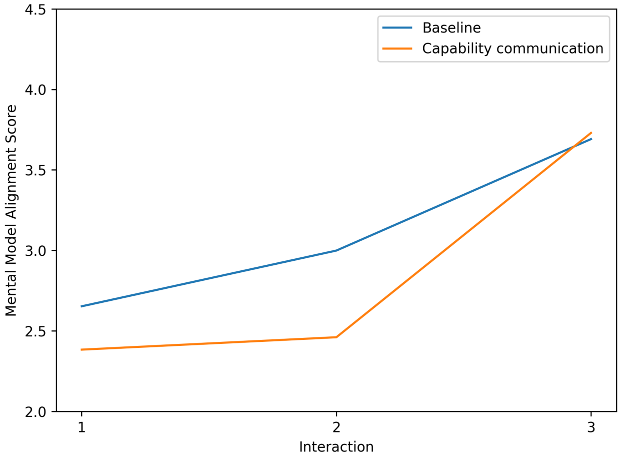

Figure 3: The capability alignment (as specified in equation 1, averaged over all participants) over the three interactions.

6 Discussion

6.1 Revisiting Our Hypotheses

While Kirschthaler et al. [11] found that increased discoverability of capabilities in conversational agents improves the user experience, we did not see a significant difference in perceived conversation quality between the baseline and the capability communication agent as we were expecting in our H1. Since the mental model capability alignment score was also not significantly different, one could assume that a greater difference in alignment could still lead to differences in perceived conversation quality. A regression analysis of the perceived conversation quality rating and the mental model alignment score did indeed show significance, however, with a small R-squared, meaning that only a small part of the variance can be explained by the model. Although the mental model alignment score seems to influence the perceived conversation quality, which supports H3, there seem to be other factors that influence it as well. One influencing factor we found is the number of misunderstandings which happen because of speech recognition problems. The agents used had the same speech recognition system and for both conditions participants reported that the agent did misunderstand them.

We did find support for H2, and were able to observe a significant learning effect regarding the agent's capabilities over time. We did not only observe a greater alignment of the mental model with the actual capabilities over time (see Figure 3), but also less uncertainty (see Figure 3). This can be attributed to the user having more time to both explicitly explore and accidentally encounter the agent's capabilities, which enables them to form a more accurate mental model. This shows that the agent's behavior is an important factor for influencing the user's expectations, as depicted in Figure 1. On the other hand, repeated interaction will lead to a similar result, highlighting the importance of capability communication in one-off interactions.

6.2 Mental Model Alignment

Although we used the ratings as a binary classification (agent has the capability - agent does not have the capability) to calculate the mental model alignment score, a more detailed analysis (see Figure 4) gave insights into the ratings depending on whether a capability was encountered or not. For capabilities that were observed by the user, the ratings reflected this, by having a high number of " I know the agent can do this" ratings, followed by "I think the agent can do this". Interestingly the ratings for the capabilities that the agent does not have and the ones that the user had not encountered were different, even though the user did not see either of them. One reason is that the user could try to use the non-existent capabilities and see that they do not work, whereas a previously unseen capability would just move into the seen category. Although unsuccessful triggering of nonexistent capabilities led to an increase in "I think the agent cannot do this" it also led to more "I do not know" answers. This indicates that even if people unsuccessfully tried to use capabilities they did not necessarily assume that the capability was not there at all, but also attributed it to themselves not using the correct words or to possible misunderstandings, like P33 who wrote:

> "I asked him to read the recipe out loud and he did not read it to me. The agent had some difficulties understanding me, therefore I cannot say for certain, that the agent cannot do that."

The ratings of the "Seen" category are also reflected in the comments providing reasoning, since most participants report that they based their decision on whether they have seen a certain capability. This shows that people pay attention to what the agent can actually do and are able to adjust their mental model and behavior based on those observations.

The comments also showed why some participants still said that they knew that the agent had a capability, even though the capability was not there. This was because they reasoned based on the general difficulty of having such a capability in an agent and the fact that we asked for this capability. The last argument is something that we had considered while selecting the capabilities we asked for, which is why we included existent and non-existent capabilities to avoid enabling participants to simply guess the capabilities correctly.

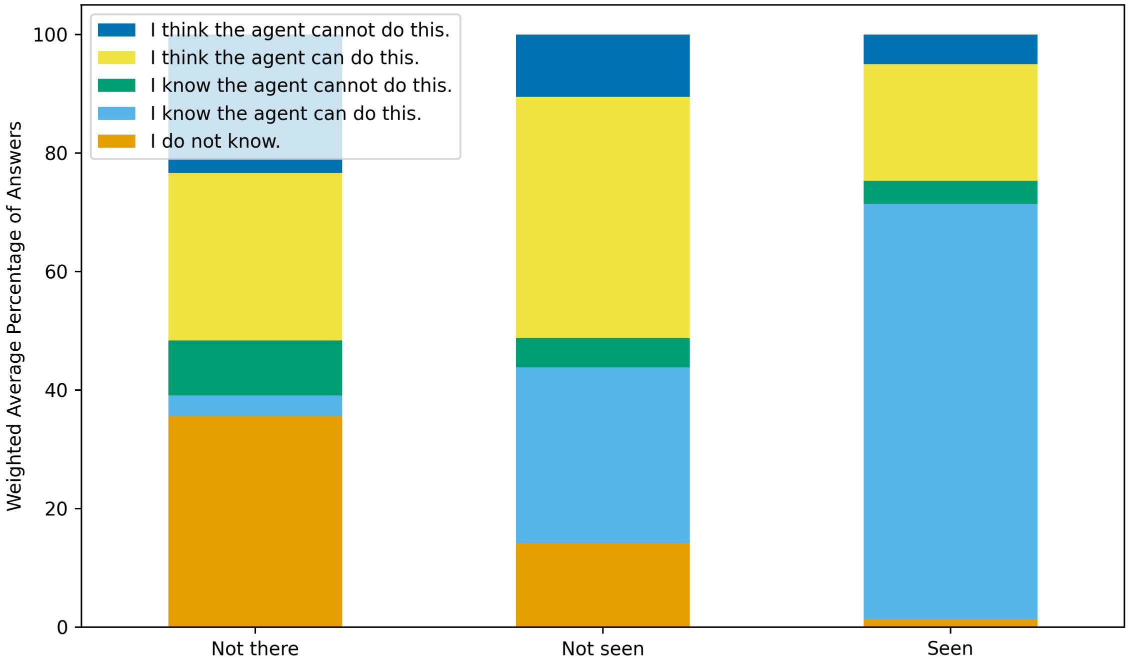

Figure 4: The user ratings for capabilities that were not there, capabilities that were there but not encountered yet and capabilities that were there and were already encountered.

However, this shows that there are also factors other than the interaction itself that shape the user's mental model.

6.3 Behavioral Effects

While there was no significant difference between the conditions regarding the perceived conversation quality, there was one for the sentence length (see Table 6). In the condition where the agent provided more information regarding its capabilities users used significantly longer sentences than in the baseline condition. The use of less total intents indicates that the users were using multiple entities of the same intent in a single sentence (e.g. "I want a recipe with *onion* and *garlic*, but no *ginger*") instead of splitting the requests into multiple sentences with the same intent. Since our agent assigned one intent per user utterance, this meant that users in the capability communication condition were using less, but longer sentences. A similar effect can be observed in human-robot interaction, where user's used longer sentences and a more conversational interaction style during interactions with a robot, if the robot communicated its capabilities proactively [21].

We found that the variation in used intents was the same in both conditions, showing that the users were still exploring to the same degree, even if they used overall less intents. Being able to complete the task, by using less utterances, while still having the same amount of variation in the used intents indicates that participants in the capability communication condition were able to use the agent more efficiently, even if they did not have more accurate explicit knowledge about the agent's capabilities. A possible reason could be that the user picks up on the capabilities of the agent without being aware of it.

6.4 Limitations

Since the capability communication agent used the conversation history to modify its utterances, the modification sometimes gave the impression that the agent understood and executed the user request, when it was not able to do it. An example for this behavior would be, if the user asks the agent for a winter recipe with carrots and the agent correctly adds carrots as a requirement but also refers to the winter requirement (which it cannot satisfy) in its response to indicate that it understood the user. This repetition of the user requests could make users believe that the agent did not only understand their speech correctly, but also followed their instructions.

While we constructed the set of capabilities used to evaluate the mental model after careful consideration, it is still only evaluating a subset of the knowledge the user has about the agent. Additionally, our approach only captures knowledge that the participants are explicitly aware of. Therefore, it might be interesting to also explore other methods of measuring the user's mental model during interactions to get a more complete picture.

The speech recognition misrecognizing or not hearing parts of utterances was a limitation in both conditions.

7 Conclusion

Comparing the interactions of users with two different agents, one that communicates its capabilities and one that uses basic repair strategies, showed that the capability communication led to a change in user behavior. Users were able to communicate more efficiently, even though they did not have more explicit knowledge about the capabilities at the level of granularity measured. For future work it would be interesting to investigate if a more fine-grained analysis would show any differences. An analysis of the mental model alignment with the actual capabilities showed a significant improvement over time, highlighting that users are able to learn about an agent's capabilities by interacting with it.

Acknowledgments

This research was (partially) funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, https://hybrid-intelligence-centre.nl, grant number 024.004.022.

References

[1] Charles R Berger and Richard J Calabrese. 1974. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human communication research* 1, 2 (1974), 99–112.

[2] Simone Borsci, Alessio Malizia, Martin Schmettow, Frank Van Der Velde, Gunay Tariverdiyeva, Divyaa Balaji, and Alan Chamberlain. 2022. The Chatbot usability scale: The design and pilot of a usability scale for interaction with AI-based conversational agents. *Personal and ubiquitous computing* 26, 1 (2022), 95–119.

[3] Elizabeth Cha, Anca D Dragan, and Siddhartha S Srinivasa. 2015. Perceived robot capability. In *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 541–548.

[4] Eric Corbett and Astrid Weber. 2016. What can I say? addressing user experience challenges of a mobile voice user interface for accessibility. In *Proceedings of the 18th international conference on human-computer interaction with mobile devices and services*. 72–82.

[5] Nathaniel Dennler, Changxiao Ruan, Jessica Hadiwijoyo, Brenna Chen, Stefanos Nikolaidis, and Maja Matarić. 2023. Design metaphors for understanding user expectations of socially interactive robot embodiments. *ACM Transactions on Human-Robot Interaction* 12, 2 (2023), 1–41.

[6] Rahul R Divekar, Jaimie Drozdal, Yalun Zhou, Ziyi Song, David Allen, Robert Rouhani, Rui Zhao, Shuyue Zheng, Lilit Balagyozyan, and Hui Su. 2018. Interaction challenges in AI equipped environments built to teach foreign languages through dialogue and task-completion. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 597–609.

[7] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[8] G Mark Grimes, Ryan M Schuetzler, and Justin Scott Giboney. 2021. Mental models and expectation violations in conversational AI interactions. *Decision Support Systems* 144 (2021), 113515.

[9] Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

[10] Laurent Karsenty. 2002. Shifting the design philosophy of spoken natural language dialogue: From invisible to transparent systems. *International Journal of Speech Technology* 5 (2002), 147–157.

[11] Philipp Kirschthaler, Martin Porcheron, and Joel E Fischer. 2020. What can i say? effects of discoverability in vuis on task performance and user experience. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–9.

[12] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the sigchi conference on human factors in computing systems*. 1–10.

[13] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 3–10.

[14] Laura Kunold, Nikolai Bock, and Astrid Rosenthal-von der Pütten. 2023. Not All Robots Are Evaluated Equally: The Impact of Morphological Features on Robots' Assessment through Capability Attributions. *ACM Transactions on Human-Robot Interaction* 12, 1 (2023), 1–31.

[15] J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* (1977), 363–374.

[16] Bingjie Liu. 2021. In AI we trust? Effects of agency locus and transparency on uncertainty reduction in human–AI interaction. *Journal of computer-mediated communication* 26, 6 (2021), 384–402.

[17] Bertram F Malle, K Fischer, J Young, A Moon, and E Collins. 2020. Trust and the discrepancy between expectations and actual capabilities. *Human-robot interaction: Control, analysis, and design* (2020), 1–23.

[18] Robert J Moore and Raphael Arar. 2019. *Conversational UX design: A practitioner's guide to the natural conversation framework*. Morgan & Claypool.

[19] Roger K Moore. 2017. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. *Dialogues with Social Robots: Enablements, Analyses, and Evaluation* (2017), 281–291.

[20] Isabela Motta and Manuela Quaresma. 2023. Increasing Transparency to Design Inclusive Conversational Agents (CAs): Perspectives and Open Issues. In *Proceedings of the 5th International Conference on Conversational User Interfaces*. 1–4.

[21] Merle M. Reimann, Koen V. Hindriks, Florian A. Kunneman, Catharine Oertel, Gabriel Skantze, and Iolanda Leite. 2025. What Can You Say to a Robot? Capability Communication Leads to More Natural Conversations. arXiv:2502.01448 [cs.RO] https://arxiv.org/abs/2502.01448

[22] Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. 2019. Considerations on explainable AI and users' mental models. In *CHI 2019 Workshop: Where is the human? Bridging the gap between AI and HCI*. Association for Computing Machinery, Inc.

[23] Francisco AM Valério, Tatiane G Guimarães, Raquel O Prates, and Heloisa Candello. 2017. Here's what I can do: Chatbots' strategies to convey their features to users. In *Proceedings of the xvi brazilian symposium on human factors in computing systems*. 1–10.

[24] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2021. Explainable embodied agents through social cues: a review. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 3 (2021), 1–24.

[25] Nicole Yankelovich. 1996. How do users know what to say? *interactions* 3, 6 (1996), 32–43.

Appendix

Table 7: The different prompts used. The "Normal" prompt is used for both conditions, the other prompts are used in the capability communication one.

Context	Prompt
Normal	"You are a part of a recipe recommendation agent. Your responsibility is to make the utterance provided by the agent sound more natural, while also taking the conversation history into account. You cannot invent new content or recipes and are only supposed to modify the utterance given to you. You get the conversation history, including all previous system and user utterances. Use the conversation history as context. You also get the current utterance, selected by another part of the recipe recommendation agent, which you are supposed to modify. Keep the modified sentence concise and relevant and do not invent anything that is not in the history or utterance. Return just the modified utterance."
Mismatch	"You are a part of a recipe recommendation agent. Your responsibility is to make the utterance provided by the agent sound more natural, while also providing information about the problem encountered. You cannot invent new content or recipes. The current problem is that the user is trying to do something that the agent cannot do at the current point of the conversation. You get the conversation history, including all previous system and user utterances, from which you can see what the user was trying to do. Use the conversation history as context for making your reply as helpful as possible. You also get the current utterance, selected by another part of the recipe recommendation agent, which you are supposed to modify using the conversation history. Keep the modified sentence concise and relevant, but directly reference the conversation history if possible, and do not invent anything that is not in the history or utterance. Return just the modified utterance."
Misunderstanding	"You are a part of a recipe recommendation agent. Your responsibility is to provide information about the problem encountered. You cannot invent new content or recipes. The current problem is that the user's utterance was not understood by the agent. You get the conversation history, including all previous system and user utterances, from which you can see what the user said that was not understood. Use the conversation history and the agent's capabilities as context for making your reply as helpful as possible. The agent can: filter on the in-exclusion of ingredients and ingredient types, dietary restrictions, cuisine, duration, difficulty, number of ingredients, number of steps, servings, tags and meal types. It can also remove filters (by value or filter name), restart, display the recipes if there are less than 100 left and can give (random) recommendations. It can display the full recipe the user selects and it can explain what it can do. It cannot do anything else. Directly reference the last user utterance and use the knowledge about the agent. You also get the current utterance, selected by another part of the recipe recommendation agent, which you are supposed to replace. Keep the new sentence concise and relevant, and do not invent anything that is not in the history or capabilities. Return just the new utterance."
User initiated	"You are a part of a recipe recommendation agent. Your responsibility is to explain what the agent can do, while also taking the conversation history into account. You cannot invent new content or recipes and are only supposed to modify the utterance given to you. The agent can filter on the in-exclusion of ingredients and ingredient types, dietary restrictions, cuisine, duration, difficulty, number of ingredients, number of steps, servings, tags and meal types. The agent can also remove filters (by value or filter name). The agent can restart, display the recipes if there are less than 100 left and can give (random) recommendations. It can display the full recipe the user selects and it can explain what it can do. It cannot do anything else. You get the conversation history, including all previous system and user utterances. Use the conversation history as context to identify what you should disclose about the agent. You also get the current utterance, selected by another part of the recipe recommendation agent, which you are supposed to modify. Keep the modified sentence concise and relevant and do not invent anything that is not in the list of capabilities provided. Return just the modified utterance."
Inefficiency	"You are a part of a recipe recommendation agent. Your responsibility is to make the utterance provided by the agent sound more natural, while also taking the conversation history into account. You cannot invent new content or recipes and are only supposed to modify the utterance given to you. You get the conversation history, including all previous system and user utterances. Use the conversation history as context. You also get the current utterance, selected by another part of the recipe recommendation agent, which you are supposed to modify. In the modified utterance also inform the user that they can add and remove multiple preferences at once. Keep the modified sentence concise and relevant and do not invent anything that is not in the history or utterance. Return just the modified utterance."