

“It’s our fault!”: insights into users’ understanding and interaction with an explanatory collaborative dialog system

Katharina Weitz, Lindsey Vanderlyn, Ngoc Thang Vu, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Weitz, Katharina, Lindsey Vanderlyn, Ngoc Thang Vu, and Elisabeth André. 2021. “It’s our fault!”: insights into users’ understanding and interaction with an explanatory collaborative dialog system.” In *Proceedings of the 25th Conference on Computational Natural Language Learning, November 2021, online*, edited by Arianna Bisazza and Omri Abend, 2021.conll-1.1. Association for Computational Linguistics.
<https://aclanthology.org/2021.conll-1.1/>.

“It’s our fault!”: Insights Into Users’ Understanding and Interaction With an Explanatory Collaborative Dialog System

Katharina Weitz*

University of Augsburg

katharina.weitz@uni-a.de

Lindsey Vanderlyn*

University of Stuttgart

vanderly@ims.uni-stuttgart.de

Ngoc Thang Vu

University of Stuttgart

thangvu.ims.uni-stuttgart.de

Elisabeth André

University of Augsburg

elisabeth.andre@uni-a.de

Abstract

Human-AI collaboration, a long standing goal in AI, refers to a partnership where a human and artificial intelligence work together towards a shared goal. Collaborative dialog allows human-AI teams to communicate and leverage strengths from both partners. To design collaborative dialog systems, it is important to understand what mental models users form about their AI-dialog partners, however, how users perceive these systems is not fully understood. In this study, we designed a novel, collaborative, communication-based puzzle game and explanatory dialog system. We created a public corpus from 117 conversations and post-surveys and used this to analyze what mental models users formed. Key takeaways include: Even when users were not engaged in the game, they perceived the AI-dialog partner as intelligent and likeable, implying they saw it as a partner separate from the game. This was further supported by users often overestimating the system’s abilities and projecting human-like attributes which led to miscommunications. We conclude that creating shared mental models between users and AI systems is important to achieving successful dialogs. We propose that our insights on mental models and miscommunication, the game, and our corpus provide useful tools for designing collaborative dialog systems.

1 Introduction

Task-oriented dialog systems are designed to help users accomplish a specific goal, such as getting a recommendation or booking an appointment. However rather than interacting as equals, the user maintains control in the dialog, all decisions passing through them. While appropriate for tasks, such as recommender systems, this form of communication leaves the labor of decisions solely on the human and discounts advantages in an Artificial

Intelligence’s (AI) reasoning ability (Grosz, 1996) which can limit the potential of human-AI partnerships (Sklar and Azhar, 2015). Collaborative dialog offers an alternative (Nothdurft et al., 2017; Narayan-Chen et al., 2019) where both parties contribute to the final strategy.

It has been a long standing goal in AI to construct computer systems which can effectively act as partners in collaborative activity (Grosz et al., 1999). As an example, if a human and robot are working together, collaborative dialog could allow the robot to communicate the scope of its abilities, enabling both partners to work together to develop an effective strategy (Nikolaidis et al., 2018). In more complex scenarios, collaborative dialog could be useful for applications, such as search and rescue (Burke et al., 2004) or medical diagnosis recommendations (Lundberg et al., 2018).

One important aspect of human-AI collaboration is how well users can form sound mental models of their partner. Put simply, a mental model is a person’s cognitive representation for how/why a complex system (e.g., a dialog system) works (Johnson-Laird, 1980; Norman, 2014; Halasz and Moran, 1983). Rather than trying to process all aspects of such systems at once, users will create a simplified model in their mind (Clements, 2004). These models can range from simple, e.g., "If I say something, the system will say something back" to arbitrarily complex e.g., "the system recognizes that if I take too long to answer then I am upset, so it tries to comfort me." These cognitive representations are shaped through both the users’ observations and previous experiences. Explainable Artificial Intelligence (XAI) aims at supporting humans to know when, whether, and why to trust and rely on an AI system (or not) (Hoffman et al., 2018; Adadi and Berrada, 2018), and therefore helps to build sound mental models. In cooperative tasks, sound mental models can help a user predict the behavior of their partner (Bansal et al., 2019) and

*Both authors contributed equally to this research.

avoid miscommunication (Doyle and Paton, 2017).

However, currently it is unclear how users perceive collaborative dialog systems and what mental models they form about them. Concretely, we then ask the following research questions:

- **RQ1. Mental Models:** (1a) How do participants perceive an AI dialog partner in a cooperative setting (e.g., trust, game engagement)? (1b) What types of mental models do participants form about an interactive AI dialog partner in a cooperative game?
- **RQ2. Miscommunication:** (2a) Do participants' attributes (e.g., age, gender, previous game experience) have an impact on their ability to successfully collaborate with an AI dialog partner? (2b) Which strategies do participants use to resolve miscommunication situations?

To answer these questions, we first created a novel cooperative puzzle game and explanatory dialog system. We then conducted an experiment with 129 participants, collected logs of their conversations and survey responses of their impressions of their dialog partner. We used these to analyze what mental models they formed, how they navigated miscommunications, and the implications these models have for dialog system design. The corpus we collected is publicly available¹.

The main contributions of this work are 1) We provide a novel combination of two data sources: (A) Self-reports (quantitative & qualitative data), and (B) behavioral data from the dialogs to gain new insights on users' mental models about an interactive AI dialog partner. 2) We demonstrate that even when users were not engaged in the task, they perceived the AI-dialog partner as intelligent and likable, implying they saw it as a partner separate from the game. 3) We show users were correctly able to identify the system's abilities, despite this they projected human attributes and motivations onto the system, leading to miscommunications 4) We find user attributes, such as age and previous experience, influenced how they interact with the system. 5) We show that despite users having a sound mental model of the dialog system and employing a variety of strategies to resolve miscommunications, they were largely unable to meaningfully resolve

them, implying a need for shared mental models between users and AI systems.

2 Related Work

2.1 Mental Models in Collaborative Dialog

During interactions with a system, users' mental models can be formed or changed (Rutjes et al., 2019). Here, XAI can support users in creating correct mental models about a system. In addition, explanations have the potential to increase trust and acceptance (Cramer et al., 2008).

Although there has been increasing interest in collaborative dialog systems (Kim et al., 2019; Polyak et al., 2017; Danise and Striegnitz, 2012; van Waveren et al., 2019), to our knowledge there has only been one experiment which sought to understand how users form mental models in this setting. In their work, Gero et al. (2020) develop a word guessing game, where the human player must guess which word their partner is thinking of based off of one word clues over multiple games.

In our study, we focus on mental models formed during a single longer interaction, rather than multiple short ones, and how a more complex dialog interaction affects the models formed.

2.2 Miscommunication in Human-AI Dialog

Miscommunication is a natural part of dialog (Healey et al., 2018), often taking the form of either non-understandings – when one party cannot parse what the other says – or misunderstandings – where one party incorrectly interprets what the other is saying (Ardissono et al., 1998; Raux et al., 2005). While there has been interesting work studying strategies in human-human dialog (Skantze, 2005) and identifying miscommunications in spoken dialog with conversational agents (Smith, 1997; Shin et al., 2002; Opfermann and Pitsch, 2017), little work has been done in the context of miscommunication and mental models in text-based dialogs.

Therefore, we seek to better understand the link between mental models and miscommunications and how user attributes might affect this.

3 Implementation

3.1 Game Design

To study collaborative dialog, we created a communication based puzzle game and a dialog system. In this game, each user is shown four puzzles, but not given instructions to solve them. Conversely, the dialog system has instructions for solving the puzzles

¹https://github.com/DigitalPhonetics/Our_Fault

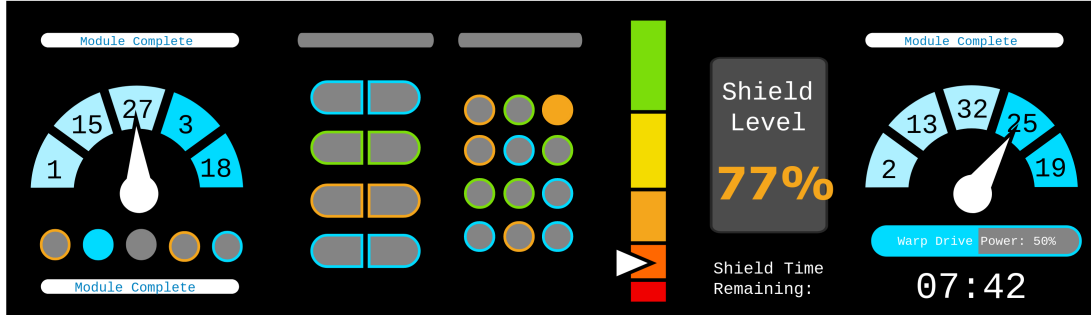


Figure 1: Game interface. There are four puzzles: The **dials module** (dials; far left and right), the **button sequence module** (row of five buttons; bottom left), the **switches module** (stack of four toggle switches; center-left), and the **button array module** (3 x 4 array of buttons; center). The shield level and timer display users’ remaining time. The warp drive indicator and indicators above each module show number of completed puzzles.

but cannot “see” or interact with them. Therefore the user must communicate with the dialog system via a text-chat interface to decide which actions to take (i.e., where to click to solve the puzzle). The game’s theme was inspired by space travel, where increasing human-AI collaboration (Bluthmann et al., 2003) is taking place. The design of the game was inspired by a cooperative control room simulator from the German Aerospace Center (Schulze Kissing and Bruder, 2016), by the human-AI collaborative games developed by Polyak et al. (2017) and van Waveren et al. (2019), and by the virtual reality game *Keep Talking and Nobody Explodes* from Steel Crate Games². The game, is comprised of a control panel (Figure 1) and a chat interface (not shown) to communicate with the dialog system.

3.2 Puzzle Modules

Three of the four puzzle modules rely on the user to give a correct description of the board (e.g., number of dials or color/sequence of buttons), the fourth, a memory game, additionally requires the user to remember their previous actions. This choice was inspired by van Waveren et al. (2019) who stress the importance of human players having a role beyond just following instructions. To simulate a spoken interaction, only the most recent dialog turn is displayed to the user. Additionally, to encourage the user to communicate with the system, there is a time limit with a time penalty for every incorrect action. To learn the game, we provided a mandatory tutorial introducing each type of puzzle³.

²<https://keeptalkinggame.com>

³Tutorial and game puzzle configurations were different

3.3 Dialog System

We used the open-source ADVISER toolkit (Li et al., 2020) to implement a rules-based conversational agent. Following the work of Cohen (2020), a rules-based system was preferred for this study as we wanted to make sure the conversational agent would behave consistently for each user. As previous work (Yin et al., 2019; Yu et al., 2017, 2016; Salem et al., 2015) has shown that trust is most damaged where mistakes are made, we hard-coded a scenario where the system made an incorrect assumption into three of the four puzzles.

The dialog system itself consists of a regex based natural language understanding module, which matches user utterances against a series of hand-crafted regexes to determine user intent. The dialog history is then stored in a Beliefstate Tracker and this information is used by the rules-based policy to decide on the next system response. Finally the system response is communicated to the user through a template based Natural Language Generation module, which chooses one of a set of pre-defined templates to convert the system intent into natural language output.

3.4 Explanation Content

For the explanations, we created a new ‘explain’ dialog act and extended the dialog policy to also give an explanation with every instruction to the user. Explanations took the form of the relevant evidence from the belief state which supported the given instruction. For example: “since the previous switch is blue, you should activate the left side”. This approach was chosen as previous work (Garcia et al., 2018; Kulesza et al., 2013) has found users best understand explanations which provide evidence rather than try to explain a process.

4 Pilot Studies

4.1 First Pilot Study

The first (Wizard-of-Oz) study was intended to verify the game concept and study design were comprehensible to users. Participants completed an introduction and a tutorial then started the main game. The dialog system was played by an experimenter, using pre-defined dialog fragments, from the dialog system. After participants played the game, they completed an online post-questionnaire⁴.

9 participants (3 female, 6 male) between the ages of 25 and 34 took part in the study. In general the game was well-accepted. Evaluation of the single game modules showed that the button array module - the module relying on users to remember their previous actions - and the switches module were most popular. Based on the free-form feedback, we exaggerated differences between similar looking elements for the main study. In general, explanations given by the dialog system were rated above average, however some participants also worried about their length (e.g., "If time were short, the explanations would have wasted some time."). Based on this feedback, explanations were made more compact for the final dialog system.

4.2 Second Pilot Study

The goal of the second study was to verify that the dialog system and user interface worked correctly and that the changes from the first pilot study had been successfully implemented. This pilot consisted of 20 participants and was performed iteratively in batches of 5 to 10 using the Amazon Mechanical Turk platform. Each participant was required to complete the game tutorial, play the game with the dialog system, and fill in a survey. On average the experience took 30 minutes and participants were compensated \$4. Based on the log files collected, we were able to identify and improve weaknesses in our natural language understanding and in the user interface.

5 Main Study

For the main study 129 participants were recruited from Amazon Mechanical Turk, with the same requirements and payment structure.

⁴Participants were in audio contact with the experimenter throughout, in case they had any problems or questions.

5.1 Evaluation Method

To understand their backgrounds, we asked participants to provide information on their age, gender, and cooperative game and video game experience.

Trust. To evaluate trust in the presented dialog system, we used the Trust in Automation (TiA) questionnaire proposed by Körber (2018), consisting of 19 items (1 = strongly disagree to 5 = strongly agree) on six subscales (i.e., reliability, understanding, familiarity, intention of developers, propensity to trust, and trust in automation).

Game Engagement. To measure the participants' engagement in the cooperative game, we used the subscales *challenge*, *negative affect*, *positive affect*, *tension*, *competence* and *flow* of the Game Engagement Questionnaire (GEQ) from IJsselstein et al. (2013).

System- and Self-Perception. To get an impression of participants' perception of the dialog partner and their own emotional state, we used the Godspeed questionnaire (Bartneck et al., 2009). The Godspeed questionnaire contains 19 semantic differentials (e.g., fake-natural; 5-point Likert scale) on five subscales (i.e., anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety). We used four of the subscales, excluding animacy.

To qualitatively evaluate participants' impressions of their dialog partner, we used the retrospection technique proposed by Hoffman et al. (2018). Participants were asked four questions about their mental model of their partner after completing the game: 1) "Describe your dialog partner (e.g., behavior, skills, impression)" 2) "What information was important for your dialog partner?" 3+4) "What was easy/difficult for your dialog partner?"⁵

5.2 Participants

129 English-speaking participants were recruited. After removing those who did not meet study criteria (e.g., bad-faith participation, duplicate participation), 117 participants (36 female, 80 male, 1 non-binary) between the ages of 21 and 64 remained. No participants had color vision impairments.

114 of the participants had heard about the term "AI", but only 8 had heard about the term "XAI". After giving the participants a definition of AI and XAI to ensure they understood the terms, they were

⁵The complete study questionnaire can be found in appendix A

asked to rate their attitude toward each on a scale ranging from 1=extremely negative to 7=extremely positive. The attitude towards AI was rated with $M = 4.99$, higher than average. The attitude towards XAI was $M = 5.88$, even higher.

Participants had greater prior experience with video games ($M = 5.94$, $SD = 1.21$) compared to cooperative games ($M = 3.97$, $SD = 1.82$), measured on a 7-point Likert scale.

6 Dataset

In this study, we collected a dialog corpus for investigating mental models and miscommunication in cooperative settings. The dataset, survey responses, analysis, and code are publicly available⁶.

Characteristics The collected corpus includes 117 dialogs with 8295 turns (system + user turns), averaging 70.9 turns per dialog, as well as user survey responses. All participant IDs were randomly assigned to ensure participant anonymity.

Annotations Each dialog consists of system and user utterances, as well as automatically generated User Act and System Act annotations. We also provide labels for the number of puzzle modules successfully completed, the number attempted, the number of mistakes the user made, and the number of mistakes the system made per dialog. Additionally, we label the miscommunication events, per dialog, with the user ID, the start and end line, which party began the miscommunication, whether the event was an example of misunderstanding or a non-understanding, and the resolution strategy which ended the miscommunication. As a note, only one strategy is recorded per miscommunication although other unsuccessful strategies may have been employed during the event. More information on the annotations is provided with the dataset. In addition to the dialogs, we also provide survey responses and annotations of content analysis labels from each user’s free-text responses.

Example Dialog An example of a successful dialog snippet from our corpus is shown below. The user and system are working on the third puzzle⁷.

SYSTEM: Okay, the switch sequence so far should be: blue, green, amber, and the slider should be set to green. Therefore,

for the second one, activate the right side of this switch.

USER: didn’t work

SYSTEM: Are you sure that the color sequence is blue, green, amber, and the slider color is green?

USER: what is the slider

SYSTEM: The Slider is a vertical bar with five sections of different size and color. The color sequence from top to bottom is green, yellow, amber, orange, and red. A triangular pointer should be sitting in one of the sections. This indicates your current slider level (which is one of the five colors). Do you see it?

USER: the slider is orange

7 Results

7.1 RQ1. Mental Model

Our first research questions were (1a) how do participants perceive an interactive AI dialog partner? And (1b) what type of mental models participants form about it in a cooperative game? To evaluate these questions, we analysed the quantitative data and the free-form feedback from our mental model questions (see section 5.1 for more details). For the quantitative data, we used one-sample t-tests⁸. In addition, we prepared the qualitative data using the conventional content analysis approach proposed by Hsieh and Shannon (2005)⁹. This was performed by two independent raters, not involved in the study design or informed of the research questions.

7.1.1 Quantitative Analysis

When looking at the quantitative data, we found participants rated their dialog partner trustworthy ($M = 2.88$, $SD = 0.77$) but did not feel strongly engaged to the game ($M = 1.58$, $SD = 0.52$) (see Table 1 on the next page).

For the Godspeed questionnaire (see Figure 2 on the following page) we found that participants perceived their dialog partner as intelligent, likeable,

⁸We used one-sample t-tests to compare the empirical values to the mean values of the used questionnaires

⁹In content analysis, user utterances are annotated from preset labels generated from the collected data. This allows trends among users to be identified and quantified, while providing users greater freedom of expression than a Likert scale.

⁶https://github.com/DigitalPhonetics/Our_Fault

⁷See appendix B for examples of miscommunication.

Table 1: In the one-sample t-tests, *significant differences were found regarding trust, game engagement (GEQ), and all used scales of the Godspeed questionnaire except anthropomorphism (no significant effect).

Measurement	$t(116)$	p	d
Trust	5.37	<.001*	0.50
GEQ			
overall	-19.01	<.001*	1.76
frustration	-4.98	<.001*	0.46
Godspeed			
intelligence	8.50	<.001*	0.79
likeability	11.85	<.001*	1.10
safety	8.11	<.001*	0.75
calm	7.23	<.001*	0.67
anthropomorphism	0.45	.652	-

d = effect size¹⁰

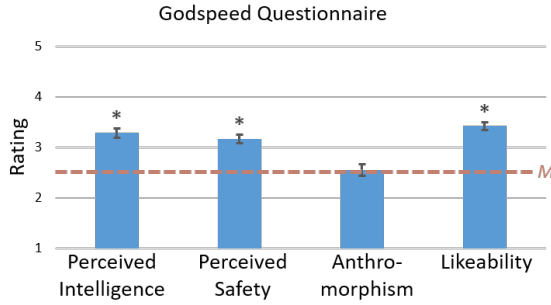


Figure 2: Results of the Godspeed Questionnaire (Bartneck et al., 2009) revealed that participants perceived their AI dialog partner as intelligent, likable, safe, but only slightly anthropomorphic. M refers to the mean value, error bars represent the standard error. * $p < .001$.

and safe, but only tended to slightly anthropomorphize it (see Table 1).

It is interesting that the participants rated their emotional state as calm (Godspeed item, $M = 3.37$, $SD = 1.30$) and not frustrated (GEQ item, $M = 1.86$, $SD = 1.38$) despite losing the game (see Table 1).

7.1.2 Qualitative Analysis

Impression of the Dialog Partner Most often participants mentioned aspects of the dialog partner’s cognitive abilities and attributes. They also often reported the dialog partner’s impact on them and the interaction quality. For each category, we report the most common subcategories (see Table 2

¹⁰Interpretation of the effect sizes according to Cohen (2013): $r = .20$ (small effect), $r = .50$ (medium effect), $r = .80$ (large effect).

on the next page for examples).

Additionally, several statements referred to human-like description of the dialog partner, including positive (e.g., “polite” or “calm”) and negative attributes (e.g., “grumpy” or “frustrated”).

Important for the Dialog Partner When looking at the question “What information was important for your dialog partner?”, we found two categories 1) Elements and 2) Communication.

Elements. Here we found four subcategories: 1) Properties of Game Elements: e.g., color, position, or on/off state. 2) Arrangement of Game Elements: e.g., sequence or number of elements in a puzzle 3) Other: e.g., what player is working on. 4) Incorrect Answers: wrong or unspecific aspects (e.g., “everything” or “don’t know”).

Communication. Here we also found four subcategories (three correct mental models and one incorrect): 1) Information: important to give the dialog partner information about the modules. 2) Feedback: giving the dialog partner feedback if things worked/ did not work. 3) How to give information: e.g., that things have to be said in a specific way 4) Incorrect Answers: e.g., that it was important for the dialog partner to write down yes/no.

Overall, 94.02% of participants formed a correct model about what was important for the dialog partner. We also note, many participants learned that not only *what* but also *how* information was communicated was important.

Easy/Difficult for the Dialog Partner The last two questions we asked participants were about their mental models of what was difficult or easy for the dialog partner. Overall, participants described 130 aspects which were easy for the dialog partner and 161 aspects that were not. Most often simple aspects of the interaction were mentioned as easy (e.g., “giving instructions”), whereas most participants addressed higher cognitive abilities and complex interaction aspects as difficult for the dialog partner (e.g., “empathetic understanding of the player” or “understanding the situation”).

7.2 RQ2. Miscommunication

Our second research questions were (2a) Do participants’ attributes have an impact on their ability to successfully collaborate with an AI dialog partner and (2b) Which strategies do participants use to resolve miscommunication situations?

Table 2: Main and sub-categories resulted from content analysis. For every sub-category (highlighted in bold), an example of participants’ free-form feedback is given. Every example response is from a different participant.

A1	They seemed to use natural language and seemed to mostly know what they were doing. (competent)
A2	They were very nice but, they had no idea of what they were talking about. (incompetent)
A3	Whenever I said something it didn’t really fit what it was looking for I suppose. (limited/simple)
B1	My dialog partner was knowledgeable and cooperative. (helpful/cooperative)
B2	I felt that the behavior in terms of response and actions was very human like. (human-like)
B3	Was an AI that was programmed to do one thing correctly. It was good at that. (machine-like)
B4	They acted cold and did not try to help much. (unhelpful)
B5	It was confused when I asked things that did not seem in the realm of its knowledge. (confused)
C1	I was impressed with how it spoke to me. (positive)
C2	Frustrating, the information they gave wasn’t descriptive enough. (negative)
D1	The AI kept on repeating the same questions when we were stuck. (poor dialog skills)
D2	It was clear in telling me what to do to solve the puzzle. (good dialog skills)

A = Dialog partners’ cognitive abilities, B = Dialog partners’ attributes,
C = Dialog partners’ impact on user, D = Dialog partners’ interaction with user.

Miscommunication Events All 117 participants lost the game. When asked afterwards, “Who made mistakes during the game?”, 58.1% of the participants stated, that both, they and the dialog partner made mistakes. 28.2% reported that only they were responsible for the mistakes during the game, 12.8% said it was alone the fault of the AI, and only one person stated that no one made a mistake during the game. Overall, we found an average dialog length of $M = 70.9$ turns ($SD = 37.5$), from which $M = 13.6$ ($SD = 20.3$) of the turns included a miscommunication (19%).

To gain a better idea of how miscommunications were spread throughout participants, we divided them into three groups based on the number of turns spent on miscommunications. We found 29 users had no turns of miscommunication, 48 had few to average (1-14) turns, and 40 had a greater than average (15+) turns of miscommunication. This indicates miscommunications were not limited to only a small subset of users. Additionally, we found a significant positive relationship¹¹ ($r_s = .38$, $p < .001$) between dialog length and the number of turns of miscommunication, showing miscommunications led to longer, less efficient dialogs. We also found a positive correlation between the number of times users misunderstood the system and vice versa ($r_s = .69$, $p < .001$), indicating miscommunications led to further miscommunications.

User Attributes To investigate what caused these miscommunications, we looked at the re-

lationships between participants’ attributes (e.g., experience with video games) and miscommunication events. We found that participants with greater experience in cooperative games were significantly more often misunderstood by the dialog system ($r_s = .19$, $p = .04$). We also found a negative relationship between user age and dialog length ($r_s = -.27$, $p = .003$) as well as a negative relationship between user age and the frequency the user misunderstood the system ($r_s = -.21$, $p = .02$). However, the reported values only reflect correlations, not causal relationships.

Miscommunication Resolution Strategies

Overall, we found 167 successful strategies belonging to nine categories (see Figure 3 on the following page). The most successful user strategy was to rephrase their answer, the most successful system strategy was to provide the user of with either new or repeated information. Around, 20% of dialogs were not resolved. However, it is also worth noting that several “successful” strategies (restart, agree with the system, solve without help), demonstrate that the user has given up on resolving the miscommunication and is instead focused on moving on with the task/dialog.

Our results show that the majority of resolution strategies were initiated by the user. However, this is not surprising, as the AI dialog partner was only equipped with simple strategies (i.e., repeating questions or information).

¹¹To investigate relations between variables, Spearman’s rank correlation was used.

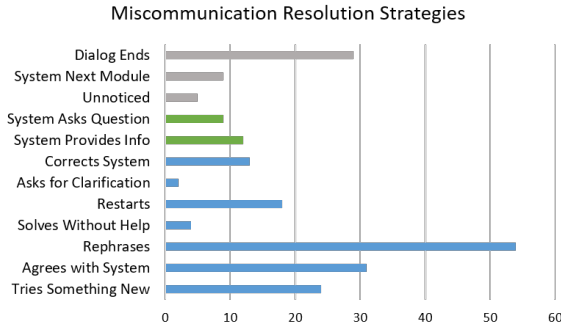


Figure 3: Strategies to resolve miscommunications and their frequency. *Blue* represents user strategies, *green* represents system strategies, and *grey* represents events that were not resolved.

8 Discussion

We gained insights into the participants’ impressions of the AI system by analyzing self-reports (quantitative & qualitative) and behavior (dialogs).

8.1 Mental Models

Quantitative Analysis Here, we wanted to know (1a) *How do participants perceive an AI dialog partner in a cooperative setting?* Our quantitative analyses show that participants were not highly engaged by the game, but perceived their AI dialog partner as intelligent and likable. In addition, our qualitative analysis showed that the dialog partner was perceived as human-like, although this was not reflected in their quantitative rating. These results suggest that the AI is seen as a separate partner playing the game with their human counterpart rather than a part of the game itself and that qualitative analysis is an important complement to quantitative.

Qualitative Analysis Here, we asked: (1b) *What types of mental models do participants form about an interactive AI dialog partner in a cooperative game?* Through interacting with the dialog system, participants correctly learned where it struggled; despite this, they expected more than it was capable of. This was evident in the fact that users attributed human characteristics to their dialog partner, which require higher cognitive abilities (e.g., intentionally misleading). In particular, participants often projected human attributes/intention to mistakes. One user stated “My impression is that I was directed in the wrong direction intentionally before being directed correctly at a time too late to correct the issues due to the time constraints.”. Similar findings were

reported by Gero et al. (2020). Here the authors found that people overestimate the AI system’s abilities, particularly those who lose the game. This suggests that it is important when designing dialog systems to transparently convey their capabilities and limitations to users. Luger and Sellen (2016) came to a similar conclusion for speech assistants, like Cortana and Siri, which suggests that our findings are transferable to other conversational agents.

8.2 Miscommunication

Miscommunication Events Although all participants lost the game, they rarely assigned fault solely to the AI in the follow-up survey, showing that users are willing to share some of the responsibility for mistakes in understanding. While many users (23) expressed that the dialog system could not completely understand them, they were often willing to accept partial blame. For example, one user described the system as, “Simple. As good as the person they’re working with.” Future work could look into what role explanation plays here as explanations are social and a part of an interactive conversation (Hilton, 1990; Miller, 2018).

User Attributes Regarding research question (2a) *Do participants’ attributes (e.g., age, gender, previous game experience) have an impact on their ability to successfully collaborate with an AI dialog partner?* we have drawn attention to the fact that users’ age and previous experiences can influence their interaction with the system. This was indicated by the significant negative correlation between participants’ cooperative game experience and the system’s misunderstandings of the user. One explanation could be that people who had prior experience with cooperative games had a mental model of how their partner should act and found it difficult that the AI partner did not match this model. Heimerl et al. (2020) reported that users tend to transfer their own mental models into AI systems for emotion recognition and expect the system to behave as they would. However, further research would be needed to confirm this for the context of cooperative games.

Miscommunication Resolution Strategies Our research question (2b) *addressed the strategies participants use to resolve miscommunication situations.* We found that miscommunications arose frequently in the cooperative dialogs, and although users employed a variety of strategies to repair

them, 20% of all events ended without being resolved and in a further 25% of cases, users gave up attempting to resolve the original miscommunication, instead only focusing on moving the dialog/task along.

This suggests that it was insufficient to rely solely on the user to repair miscommunication events. Rather, having a mechanism for resolving misunderstandings or resetting back to common ground is important. With each miscommunication event lasting an average of 7.5 turns, once a misunderstanding occurred it often precipitated a spiral, with neither party able to resolve the initial ambiguity. This limited scope of resolution was mentioned negatively by several (12) participants, e.g., “the AI kept on repeating the same questions when we were stuck”.

Our findings suggest that users can correctly identify information needed by an AI dialog partner, but this does not automatically lead to fewer miscommunications. This suggests that not only the user must build a correct mental model, but also the AI system. For human-human interactions, [Doyle and Paton \(2017\)](#) stated it is crucial that cooperating teams have a “shared mental model”. A similar demand is also made in human-AI interactions ([Gervits et al., 2020](#)).

9 Future Work

To reduce miscommunication in human-AI dialogs, it could be interesting to see the impact of more sophisticated methods (e.g., pretrained language models) and if they would increase the AI system’s ability to understand users and lead to different user-perceptions. In addition, we recommend reducing the responsibility for users. As our experiment shows, resolving miscommunication in systems like we created was mainly the responsibility of the user as the system was not able to recognize when the user had a poor understanding of the current problem/task. Therefore, future research should focus not only on promoting the user to develop correct mental models of the system, but also on how the AI system can develop a correct mental model of the human counterpart in order to promote successful human-AI cooperation.

Regarding our correlation results, it would be interesting to explore a more nuanced approach to the role of user attributes (i.e., user age, experience in cooperative games) on dialogs with an AI in a cooperative setting. In particular, it would be inter-

esting to further explore the role of user age as well as the question of how pre-existing mental models (e.g., gained through prior experience with cooperative games) promote or hinder collaboration with an AI.

In addition, a comparison with other collaborative settings could gain new insights into how task-dependent or independent the user perception of an AI system is.

10 Conclusion

In this work, we designed a novel collaborative game and dialog system to collect a new corpus and investigate how users perceive an AI dialog partner and what mental models they formed.

We found that users perceived their AI dialog partner as intelligent and likeable. Almost all users were able to identify what type of information was important to share with their partner, and many also realized this information needed to be provided in a specific way. However, although users perceived the limitations of the dialog system, they tended to overestimate its abilities and attribute human characteristics to it, resulting in miscommunication.

Users were also willing to share blame when a misunderstanding occurred, indicating they viewed their partner as sophisticated enough to be responsible for mistakes. Based on this, we consider it important for successful collaborative dialog systems that a shared (correct) mental model between user and AI system is developed. Our game, corpus, and results provide insight into users’ mental models as well as miscommunications during a dialog and serve as a tool for other researchers interested in collaborative dialog.

11 Ethical Considerations

The presented studies in this paper as well as the collected corpus have been approved by the data protection officer of the University of Augsburg. For information on the dataset collection methodology, participants, and participant compensation, please see Section 4 on page 4, Section 5 on page 4, and Section 6 on page 5.

Acknowledgments

This work was partially funded by Deutsche Forschungsgemeinschaft (DFG) under project number 442607480, PANORAMA. Ngoc Thang Vu was funded by the Carl Zeiss Foundation.

References

- Amina Adadi and Mohammed Berrada. 2018. [Peeking inside the black-box: A survey on Explainable Artificial Intelligence \(XAI\)](#). *IEEE Access*, 6:52138–52160.
- Liliana Ardissono, Guido Boella, and Rossana Damiano. 1998. A plan-based model of misunderstandings in cooperative dialogue. *International Journal of Human-Computer Studies*, 48(5):649–679.
- Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 2–11.
- Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics*, 1(1):71–81.
- William Bluethmann, Robert Ambrose, Myron Diftler, Scott Askew, Eric Huber, Michael Goza, Fredrik Rehnmark, Chris Lovchik, and Darby Magruder. 2003. Robonaut: A robot designed to work with humans in space. *Autonomous robots*, 14(2-3):179–197.
- Jennifer L Burke, Robin R Murphy, Michael D Coover, and Dawn L Riddle. 2004. Moonlight in miami: Field study of human-robot interaction in the context of an urban search and rescue disaster response training exercise. *Human-Computer Interaction*, 19(1-2):85–116.
- DH Clements. 2004. Perspective on “the child’s thought and geometry”. *Classics in mathematics education research*, pages 60–66.
- Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. Academic press.
- Philip R Cohen. 2020. Back to the future for dialogue research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13514–13519.
- Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-adapted interaction*, 18(5):455.
- Andrew Danise and Kristina Striegnitz. 2012. A collaborative puzzle game to study situated dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 8.
- Emma EH Doyle and Douglas Paton. 2017. Decision-making: preventing miscommunication and creating shared meaning between stakeholders. In *Observing the Volcano World*, pages 549–570. Springer.
- Francisco Javier Chiyah Garcia, David A Robb, Xingkun Liu, Atanas Laskov, Pedro Patron, and Helen Hastie. 2018. Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 99–108.
- Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. 2020. Mental models of ai agents in a cooperative game setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12.
- Felix Gervits, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz. 2020. Toward genuine robot teammates: Improving human-robot team performance using robot shared mental models. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 429–437.
- Barbara J Grosz. 1996. Collaborative systems (aaai-94 presidential address). *AI magazine*, 17(2):67–67.
- Barbara J Grosz, Luke Hunsberger, and Sarit Kraus. 1999. Planning and acting together. *AI magazine*, 20(4):23–23.
- Frank G Halasz and Thomas P Moran. 1983. [Mental models and problem solving in using a calculator](#). In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 212–216.
- Patrick GT Healey, Jan P De Ruiter, and Gregory J Mills. 2018. Editors’ introduction: Miscommunication. *Topics in cognitive science*, 10(2):264–278.
- A. Heimerl, K. Weitz, T. Baur, and E. Andre. 2020. [Unraveling ml models of emotion with nova: Multi-level explainable ai for non-experts](#). *IEEE Transactions on Affective Computing*, (01):1–1.
- Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin*, 107(1):65.
- Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2018. [Metrics for explainable AI: challenges and prospects](#). *CoRR*, abs/1812.04608.
- Hsiu-Fang Hsieh and Sarah E Shannon. 2005. Three approaches to qualitative content analysis. *Qualitative health research*, 15(9):1277–1288.
- Wijnand A IJsselstein, Yvonne AW de Kort, and Karolien Poels. 2013. The game experience questionnaire. *Eindhoven: Technische Universiteit Eindhoven*, pages 3–9.
- Philip N Johnson-Laird. 1980. Mental models in cognitive science. *Cognitive science*, 4(1):71–115.

- Jin-Hwa Kim, Nikita Kitaev, Xinlei Chen, Marcus Rohrbach, Byoung-Tak Zhang, Yuandong Tian, Dhruv Batra, and Devi Parikh. 2019. Codraw: Collaborative drawing as a testbed for grounded goal-driven communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6495–6513.
- Moritz Körber. 2018. Theoretical considerations and development of a questionnaire to measure trust in automation. In *Congress of the International Ergonomics Association*, pages 13–30. Springer.
- 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*, pages 3–10. IEEE.
- Chia-Yu Li, Daniel Ortega, Dirk Vāth, Florian Lux, Lindsey Vanderlyn, Maximilian Schmidt, Michael Neumann, Moritz Völkel, Pavel Denisov, Sabrina Jenne, Zorica Karacevic, and Ngoc Thang Vu. 2020. Adviser: A toolkit for developing multi-modal, multi-domain and socially-engaged conversational agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020) - System Demonstrations*. Association for Computational Linguistics.
- 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*, pages 5286–5297.
- Scott M Lundberg, Bala Nair, Monica S Vavilala, Mayumi Horibe, Michael J Eisses, Trevor Adams, David E Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim, et al. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760.
- Tim Miller. 2018. [Explanation in artificial intelligence: Insights from the social sciences](#). *Artificial Intelligence*, 267:1–38.
- Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415.
- Stefanos Nikolaidis, Minae Kwon, Jodi Forlizzi, and Siddhartha Srinivasa. 2018. Planning with verbal communication for human-robot collaboration. *ACM Transactions on Human-Robot Interaction (THRI)*, 7(3):1–21.
- Donald A Norman. 2014. [Some observations on mental models](#). In *Mental models*, pages 15–22. Psychology Press.
- Florian Nothdurft, Pascal Bercher, Gregor Behnke, and Wolfgang Minker. 2017. User involvement in collaborative decision-making dialog systems. In *Dialogues with Social Robots*, pages 129–141. Springer.
- Christiane Opfermann and Karola Pitsch. 2017. Reprompts as error handling strategy in human-agent-dialog? user responses to a system’s display of non-understanding. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 310–316. IEEE.
- Stephen T Polyak, A von Davier, and Kurt Peterschmidt. 2017. Analyzing game-based collaborative problem solving with computational psychometrics. In *Proceedings of 23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- Antoine Raux, Brian Langner, Dan Bohus, Alan W Black, and Maxine Eskenazi. 2005. Let’s go public! taking a spoken dialog system to the real world. In *Ninth European conference on speech communication and technology*.
- Heleen Rutjes, Martijn Willemsen, and Wijnand IJsselsteijn. 2019. Considerations on explainable ai and users’ mental models. In *Where is the Human? Bridging the Gap Between AI and HCI*, United States. Association for Computing Machinery, Inc.
- Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot? effects of error, task type and personality on human-robot cooperation and trust. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1–8. IEEE.
- Dirk Schulze Kissing and Carmen Bruder. 2016. Der Einsatz Synthetischer Aufgabenumgebungen zur Untersuchung kollaborativer Prozesse in Leitzentralen am Beispiel der “generic Control Center Task Environment”(ConCent). *Kognitive Systeme*, 2016(1).
- Jongho Shin, Shrikanth S Narayanan, Laurie Gerber, Abe Kazemzadeh, and Dani Byrd. 2002. Analysis of user behavior under error conditions in spoken dialogs. In *Seventh International Conference on Spoken Language Processing*.
- Gabriel Skantze. 2005. Exploring human error recovery strategies: Implications for spoken dialogue systems. *Speech Communication*, 45(3):325–341.
- Elizabeth I Sklar and Mohammad Q Azhar. 2015. Argumentation-based dialogue games for shared control in human-robot systems. *Journal of Human-Robot Interaction*, 4(3):120–148.
- Ronnie W Smith. 1997. An evaluation of strategies for selective utterance verification for spoken natural language dialog. In *Fifth Conference on Applied Natural Language Processing*, pages 41–48.

- Sanne van Waveren, Elizabeth J Carter, and Iolanda Leite. 2019. Take one for the team: The effects of error severity in collaborative tasks with social robots. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 151–158.
- Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12.
- Kun Yu, Shlomo Berkovsky, Dan Conway, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2016. Trust and reliance based on system accuracy. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, pages 223–227.
- Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 307–317.

A Survey Questionnaire from the Main Study

The following questions were asked **before** the experiment

A.1 Personal Information

What is your age? _____

To which gender identity do you most identify: ☐ male ☐ female ☐ other: _____

Do you have colour vision impairment? ☐ yes ☐ no

How do you rate your experience with cooperative games (e.g., Coop games, Escape Room games...)?

No experience ☐ ☐ ☐ ☐ ☐ ☐ ☐ Much experience

How do you rate your experience with playing video/computer games?

No experience ☐ ☐ ☐ ☐ ☐ ☐ ☐ Much experience

How well do the following statements describe your personality?

I see myself as someone who ...

	Disagree strongly				Agree strongly
...is reserved	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...is generally trusting	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...tends to be lazy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...is relaxed, handles stress well	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...has few artistic interests	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...is outgoing, sociable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...tends to find fault with others	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...does a thorough job	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...gets nervous easily	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
...has an active imagination	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

A.2 AI and XAI Knowledge

The following questions ask about Artificial Intelligence (AI). Colloquially, the term “artificial intelligence” is often used to describe machines (or computers) that mimic “cognitive” functions that humans associate with the human mind, such as “learning” and “problem solving”.

Have you ever heard about the term Artificial Intelligence (AI)? ☐ yes ☐ no

Do you have experience with Artificial Intelligence (AI)

No experience ☐ ☐ ☐ ☐ ☐ ☐ Much experience ☐

What is your general attitude towards Artificial Intelligence (AI)?

Extremely negative ☐ ☐ ☐ ☐ ☐ ☐ Extremely positive ☐

The following questions ask about Explainable Artificial Intelligence (XAI). With the help of explainable artificial intelligence (XAI) it should be easier for humans to understand the behaviour and outcome of artificial intelligence (AI).

Have you ever heard of the term Explainable Artificial Intelligence (XAI)? ☐ yes ☐ no

Do you have experience with Explainable Artificial Intelligence (XAI)

No experience ☐ ☐ ☐ ☐ ☐ Much experience ☐

What is your general attitude towards Explainable Artificial Intelligence (XAI)?

Extremely negative Extremely positive

☐ ☐ ☐ ☐ ☐ ☐ ☐

AI agents are already able to perform some complex tasks. Examples for such intelligent agents are search engines, chatbots, chessbots, and voice assistants. Suppose that AI agents would achieve high-level performance in more areas one day:

How do you expect the overall impact of such AI agents to be on humanity in the long run?

Extremely negative Extremely positive

☐ ☐ ☐ ☐ ☐ ☐ ☐

The following questions were asked **after** the experiment

A.3 Evaluation of the Game

You have played the online game Space Jam with the help of your dialog partner. In the following, we would like to know your experiences and impressions about the game.

(Game Engagement Questionnaire (GEQ) from IJsselstein et al. (2013))

Please indicate how you felt while playing the game for each of the items:

	Not at all				Extremely
I felt content	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt skilful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was interested in the game's story	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I thought it was fun	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was fully occupied with the game	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt happy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It gave me a bad mood	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I thought about other things	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found it tiresome	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt competent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I thought it was hard	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It was aesthetically pleasing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I forgot everything around me	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt good	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was fully occupied with the game	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was good at it	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt bored	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt successful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt imaginative	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt that I could explore things	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I enjoyed it	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was fast at reaching the game's targets	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt annoyed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt pressured	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt irritable	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I lost track of time	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt challenged	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I found it impressive	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was deeply concentrated in the game	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt frustrated	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It felt like a rich experience	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I lost connection with the outside world	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I felt time pressure	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I had to put a lot of effort into it	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

What did you find particularly positive about the game? What did you particularly enjoy?

What did you find particularly negative about the game? What in particular did you not enjoy?

A.4 Evaluation of the Dialog Partner

In the following we would like to collect your impression of the dialog partner and the explanations it gave you during the game.

(Sub-scales: Anthropomorphism, Likeability, and Perceived Intelligence of the Godspeed Questionnaire from Bartneck et al. (2009))

Please rate your impression of the dialog partner on these scales

Fake	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Natural
Machine-like	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Human-like
Unconscious	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Conscious
Artificial	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Lifelike
Dislike	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Like
I felt happy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Natural
Unfriendly	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Friendly
Unkind	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Kind
Unpleasant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Pleasant
Awful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Nice
Incompetent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Competent
Ignorant	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Knowledgeable
Irresponsible	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Responsible
Unintelligent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Intelligent
Foolish	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Sensible

(Sub-scale: Perceived Safety of the Godspeed Questionnaire from Bartneck et al. (2009))

Please rate your emotional state on these scales

Anxious	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Relaxed
Agitated	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Calm
Quiescent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Surprised

Describe your dialog partner (e.g., behavior, skills, impression...):

What information was important for your dialog partner?

What do you think was easy for your dialog partner?

What do you think was difficult for your dialog partner?

(Trust in Automation (TiA) questionnaire proposed by Koerber (2018))

Please rate the following statements:

	Strongly disagree	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Strongly agree
The dialog partner is capable of interpreting situations correctly	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The dialog partner state was always clear to me	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I already know similar dialog partners	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The developers are trustworthy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
One should be careful with unfamiliar automated systems	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The dialog partner works reliably	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The dialog partner reacts unpredictably	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The developers take my well-being seriously	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I trust the dialog partner	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A malfunction of the dialog partner is likely	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I was able to understand why things happened	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I rather trust a system than mistrust it	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The dialog partner is capable of taking over complicated tasks	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I can rely on the dialog partner	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
The dialog partner might make sporadic errors	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
It is difficult to identify what the dialog partner will do next	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I have already used similar systems	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Automated systems generally work well	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
I am confident about the dialog partner's capabilities	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

You completed the task. Thank you!

B Examples of User Strategies to Resolve Miscommunication in the First Module

B.1 Agrees With System

The user gives up trying to solve the puzzle and just agrees with the system.

SYSTEM: Okay, based on what you've said there should be two half-circle dials and the slider should point to green. Therefore, you should move both pointers to the middle position.

USER: i only can get one to work

SYSTEM: If I have this correct, there are two half-circle dials and the slider is set to green. So you should set both pointers to the middle position.

USER: I can only set one of the dials

SYSTEM: Alright, so there should be one dial and the slider level should be green. So you should set the pointer to the far left position.

USER: okay

B.2 Corrects System

The user informs the system there is a fourth button it didn't know about.

SYSTEM: I'm sorry, I didn't understand. Please rephrase?

USER: the button order

SYSTEM: The complete sequence of clickable (colorful border) buttons is: amber, blue, amber. Therefore, the first button should be activated.

USER: there are 4 buttons

SYSTEM: What color is the fourth clickable button? Clickable buttons have a blue, amber, or green border.

B.3 Asks for Clarification

The user realizes they don't know what the slider is and asks for clarification.

USER: What should we do next?

SYSTEM: I'm sorry, I didn't understand. Please rephrase?

USER: Hello?

SYSTEM: There are 2 half dials and the slider is set to green. Therefore, both pointers should be set to the middle position.

USER: What slider is set to green?

SYSTEM: The Slider is a vertical bar with five sections ... (truncated to save space)

B.4 Solves Without Help

The user solves the puzzle without help from the system.

SYSTEM: What color is the first clickable button? Clickable buttons have a blue, amber, or green border.

USER: yse

SYSTEM: Sorry, I didn't understand that. Could you please say it again in other words?

USER: i did that module

B.5 Restarts

The user doesn't know how to correct the system, so just restarts the module.

SYSTEM: I'm sorry, I didn't understand. Please rephrase?

USER: The sequence I gave you was wrong

SYSTEM: The complete sequence of clickable (colorful border) buttons is: amber, green, amber. Therefore, the first button should be activated.

USER: restart

B.6 Tries Something New

After ignoring system question, the user tires different phrases until their input is recognized again.

USER: lets play

SYSTEM: Sorry, I didn't understand that. Could you please say it again in other words?

USER: hello?

SYSTEM: How many dials are there on the control panel? There might be one or two of them.

B.7 Rephrases

The user gives too much information at once, rephrases to only give one piece per turn.

USER: There are two half dials, but the slider color is yellow

SYSTEM: I can't process the information this fast, please only tell me one thing at a time.

USER: the slider is yellow

SYSTEM: What number is the pointer on the left dial pointing to?