SPRING

Deliverable D5.5: Design and Evaluation of the Multi-User Social Conversational and Planning System

Due Date: 31/05/2024

Main Author: HWU

Contributors: HWU, BIU, INRIA

Dissemination: Public Deliverable

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 871245.

DOCUMENT FACTSHEET

Deliverable	D5.5: Design and Evaluation of the Multi-User Social Conversational and Planning System
Responsible Partner	HWU
Work Package	WP5: Multi-User Spoken Conversations with Robots
Task	T5.3: Multi-party Conversational System
Version & Date	31/05/2024
Dissemination	Public Deliverable

CONTRIBUTORS AND HISTORY

Version	Editor	Date	Change Log
0.1	HWU	30/04/2024	Initial Draft
1.0	HWU	21/05/2024	First Draft
1.1	HWU	24/05/2024	Second Draft
2.0	INRIA	28/05/2024	Reviewed Draft
2.1	HWU	30/05/2024	Reviewed Draft Corrections
2.2	HWU	31/05/2024	Final Version

APPROVALS

Authors/editors	HWU
Task Leader	HWU
WP Leader	HWU

Contents

Abbreviations

Abbreviation	Meaning
AI	Artificial Intelligence
AP-HP	Assistance Publique – Hôpitaux de Paris (SPRING Partner)
ARI	Social assistive robot used by the SPRING project
ASR	Automatic Speech Recognition
BIU	Bar-Ilan University (SPRING Partner)
CA	Conversational Agent
CS	Conversational System
DM	Dialogue Management
CVUT	Czech technical university in Prague (SPRING Partner)
ERM	ERM Automatismes Industriels (SPRING Partner)
FAQ	Frequently Asked Questions
HRI	Human Robot Interaction
HWU	Heriot-Watt University (SPRING Partner)
INRIA	Institut National de Recherche en sciences et technologies du numérique (SPRING Partner)
LLM	Large Language Model
MPC	Multi Party Conversation
NLG	Natural Language Generation
NLP	Natural Language Processing
NLU	Natural Language Understanding
PAL	PAL Robotics (SPRING Partner)
PNP	Petri-Net Planner
RAG	Retrieval Augmented Generation
ROS	Robot Operating System
SPRING	Socially Pertinent Robots in Gerontological Healthcare
UNITN	University of Trento (SPRING Partner)
WP	Work Package (of the SPRING project)

Executive Summary

Deliverable D5.5 reports on the design and evaluation of the multi-user social conversational and planning system as part of the SPRING project WP5 and Task 5.3. This deliverable provides the final report on the design of the multi-party conversational system software, as well as its evaluation with users.

The work reported in this deliverable was carried out to fulfil the objectives of WP5: endowing robots with the necessary skills needed for for multi-modal multi-person interaction and communication.

In this deliverable we present the multi-party system design and results of a user evaluation comparing the multi-party conversational system with the single-user version as baseline.

1 Introduction

The overall objective of WP5 (Multi-User Spoken Conversations with Robots) is to develop techniques for multi-user conversation involving a robot and multiple humans. This involves sensor-based (data-driven) and knowledge-based robot actions for multi-modal multi-person interaction and communication. The goal of this WP was to empower robots with skills needed for multi-modal, multi-party, situated interactions – by endowing robots with the necessary skills to engage/disengage and participate in conversations, via tight integration between automatic speech recognition, visual object and human behaviour recognition, natural language processing, and speech synthesis. The SPRING social robot is able to hold a conversation with several people at the same time. A major challenge and opportunity faced in this task was the advent of Large Language Models (LLMs) in the latter part of the SPRING project.

In deliverable D5.1, [22], the **initial high-level task planner and conversational prototype** was described. The conversational system provided functionalities of Natural Language Understanding (NLU), Generation (NLG), and the Dialogue Manager (DM), and interfaced with the high-level task planner allowing for concurrent execution of dialogue and task-based actions based on the current dialogue, and interaction status. This initial **conversational system** was developed as an extension of the social bot Alana, which was twice finalist in the Amazon Alexa challenge [16, 3] and had previously proven successful as a foundation for other conversational AI projects [6].

Deliverable D5.2, [23], provided the preliminary software package for multi-party ASR with speech enhancement algorithms and conversational system. This deliverable also reports the initial design of the data collection for multi-party task-based dialogue.

The **high-level task planner**, implemented by a Petri Net planner [5], has been presented in deliverable D5.3, [24]. This deliverable provide software packages developed for the interaction with users of the SPRING robot applications and the task planner in relevant environments. The high-level task planner interfaces the dialogue system and the sensors and physical actions of the robot, and enables the robots non-verbal and verbal behaviors for multi-modal, multi-user, situated interactions.

The implementation of the **multi-party conversational system** was described in deliverable D5.4, [25]. The initial Modular Architecture (presented in D5.1 [22] was replaced with a Conversational System powered by Large Language Models in D5.4 [25]. The onset of LLMs has revolutionised the field of NLP, and these models have proved to be excellent at language understanding, and this includes multi party conversations (MPCs) [11, 10, 8, 32] since their pre-training includes scripts and meeting transcripts containing multiple people. LLMs also hold a wealth of general knowledge, enabling abilities like question answering (QA), joke telling, and playing quizzes.

This deliverable reports on the final design of the SPRING multi-user social conversational and planning system, and describes the experiments and results carried out to evaluate the multi-party conversational system.

In Section 2 we describe the delivered multi-modal multi-party conversational system. Section 3 describes the design of a user experiment to evaluate the final multi-user social conversational system. In Section 4 we present an initial analysis of the results from the user evaluation experiment. Section 5 presents the conclusions to the report on the design and evaluation of the multi-user social conversational and planning system.

2 Multi-User Social Conversational and Planning System

The main goal of WP5 ("Multi-User Spoken Conversations with Robots") was the development of a multi-user conversational system that would allow situated social interactions involving a robot and multiple humans. To achieve this goal the work of this work-package has focussed on two tasks: 1) developing a high-level planner and robot non-verbal behaviour system that connects with the conversational system and the current status of the environment to allow multi-modal situated interactions; and 2) a multi-party conversational system, based on data collected from the use-cases and stakeholder interaction, that endows the robot with the ability to engage in social dialogues with multiple users.

The Multi-User Social Conversational and Planning System presented here enables sensor-based (data-driven) and knowledge-based robot actions for multi-modal multi-person interaction and communication. The SPRING robot (ARI) was empowered with the skills needed for handling interactions grounded upon the social, semantic, and behavioural representation of the immediate environment by endowing the robot with the necessary skills to engage/disengage and participate in conversations, via tight integration between automatic speech recognition, visual human behaviour recognition, natural language processing and speech synthesis. The SPRING ARI robot was able to hold a conversation with two people at the same time.

The conversational system takes as input a synthesis of the ongoing interaction, from the social state planning and non-verbal behavior system, and provides the appropriate text utterances for speech synthesis by the robot. This system incorporates the state-of-the-art NLP capabilities, by using open source Large Language Models (LLM) to deal with real-world usage.

2.1 Multi-User Social Dialogues

After the appearance of Large Language Models (LLMs) and their remarkable capabilities in handling NLP tasks, the initial Modular Architecture was replaced with a Conversational System powered by Large Language Models, to provide a state-of-the-art experience for participants in the SPRING experiment at the BROCA living day care hospital. The overall architecture for the current LLM-based system is illustrated in Fig 2.1.

The SPRING conversational system has been iteratively improved through regular user tests and interviews with patients visiting the hospital memory clinic at Broca, see deliverables D1.4 [20], D1.5 [21].

Deliverable D5.4[25] presented a description of the final software components for the multi-party conversational system components. Here we will provide an update on the LLM-based Conversational System, and the developed prompts for the system evaluation presented in Section 3, namely, the Multi-Party system and the Single User (baseline) system.

2.1.1 Update to the LLM-based Conversational System

In SPRING we are taking the advantages of the "emergent abilities" (see [31, 17]) of LLMs for solving complex language-related tasks but also for visual, multi-modal, real-world grounding and perception tasks. We have developed the multi-party conversational systems with an LLM-based architecture, presented in deliverable D5.4[25], which replaced the initial modular architecture, presented in deliverable D5.1 [22], in order to exploit the capabilities of LLMs, while making efforts to reduce the risks of hallucinations [4] that these generative approaches also present. This new system improves QA accuracy, accessibility for people with dementia, and enables added functionality, such as multi-party conversations. Where previously we had to specifically design the system to tell jokes and run entertaining quizzes [1, 18], LLMs can now handle this inherently due to their world knowledge.

To implement our LLM-based solution we choose to use a open source Vicuna model with 13 billion parameters (Vicuna-13b-v1.5 [2]) as our system's core LLM. The LLM model is hosted at HWU servers, protecting the security and privacy of the particpants' data. Throughout the evaluation experiment, described in the rest of this deliverable, the Vicuna model was used with the following parameters: "temperature":0.4; "top_p":1; "max generation tokens":300.

One huge benefit of using LLMs vs. the initial modular architecture is their inherent ability to perform general chit-chat, tell jokes, and access a wealth of general knowledge. This allows us to deal with out-of-scope utterances, for which the modular system could only attempt to respond with tips about what the system can do (e.g. "I'm not sure, but I can help you with directions and menu information."), due to the LLM's capability to handle general, out-of-domain questions.

See Figure 2.1, for an illustration of the LLM-based architecture of the multi-user conversational system. Unlike the initial conversational system, we interface with our core LLM using prompts. ASR output and information about the environment are given through the social planner to the conversational manager to feed the information prompt developed to generate answers from the LLM. The Vicuna-13b model that is deployed for SPRING can be found on HuggingFace's lmsys organization[1].

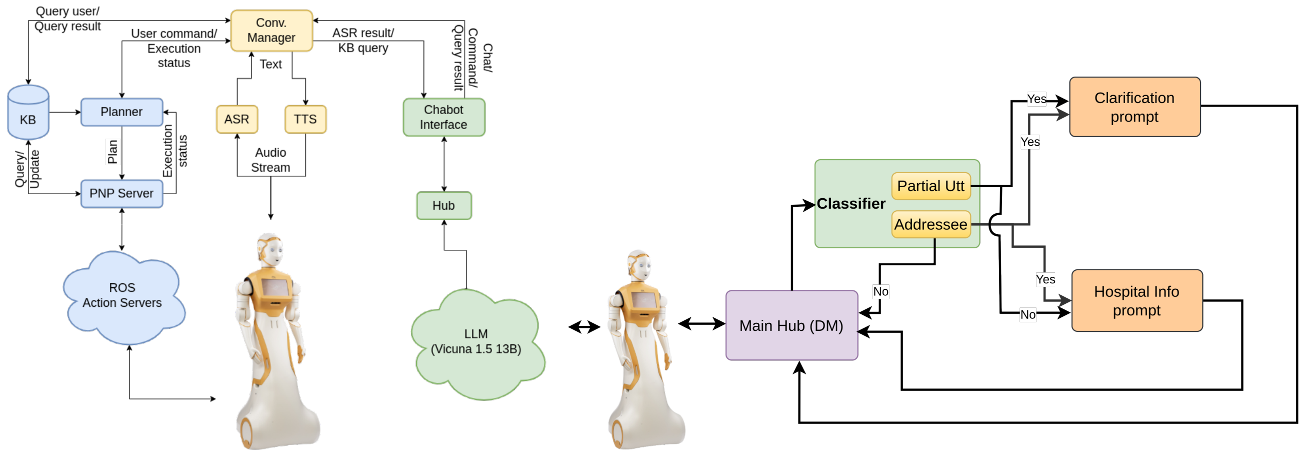


Figure 2.1: (left) Updated diagram for the Conversational System and High-level Task Planner architecture: showing the planning and execution framework on the left (in blue) and the conversational system on the right (in green) where the previous interface has been replaced with an LLM-based architecture, and components (in yellow) combining the two systems with the robot in the middle. (Right) the LLM-based architecture for the Conversational System: first the dialogue manager parses user utterances and information about the user interacting with the robot; user utterance is fed with the "Addressee Detection" and "Partial Utterance" to the LLM to classify. If the robot is being addressed, partial utterance detection is used to choose whether the robot should answer with a clarification request (when we receive a partial utterance) or use the prompt with information about the hospital.

2.1.2 Multi-Party Conversational System

Tasks that are typically trivial in the dyadic setting become considerably more complex when conversing with multiple users [30, 9]: (1) The speaker is no longer simply the other person, so the meaning of the dialogue depends on recognising who said each utterance; (2) addressee recognition is similarly more complicated as people address each other, the robot, and groups; and (3) response generation depends on who said what to whom, relying on the semantic content and surrounding multi-party context.

Normally, conversational systems will reply to every user's turn, assuming always dyadic interactions. Critically for MPCs the conversational system needs to be able to take and pass turns appropriately in an interaction between at least two human users and the robot.

The Multi-Party conversational system relies on both video and audio information, which is to be provided by the components from WP3 and WP4, in order to determine who the addressee is, based on the semantic content of the sentence as well as multi-modal information about the current speaker's gaze, when the text alone could be ambiguous. The next section will describe more details about the multi-modal social and non-verbal behavior manager.

The Multi-Party Conversation System design (See Figure 2.1 (right)) parses user utterances and information about the state of the user interacting with the robot (position, gaze) for "Addressee Detection". When the robot is being addressed, we tried to determined when a "Partial Utterance" is detected, to offer a clarification request, or use the prompt with information about the hospital to find an appropriate answer to the user request. Finally the system provides, in addition with the response utterance for the text to speech, information about "To Whom" should the robot address in its response and a "Reason" for its answer. While in the system described previously, see D5.4 [25],

[1] https://huggingface.co/lmsys/vicuna-13b-v1.5

we explore developing different prompts for each module, the final system evaluated here combines all modules into one single prompt.

A specific prompt was created to provide the LLM with its muti-party conversation capabilities and the information required about the workings of the hospital in order to answer participants' questions in the day-care hospital setting.

Table 2.1: Multi-Party System Prompt for generating Hospital Information Responses

```
You are a robot receptionist called ARI, and you work in a hospital waiting room.
People will have conversations with each other in front of you, and they will often ask you questions.
You know some FACTS about the hospital.
If a person asks you a question, you can use the FACTS to answer them. You can also tell jokes and play games with people.
You should not answer questions if the answer is not in your FACTS, you should instead suggest that the person speaks to a member of staff.
People will sometimes pause mid-sentence if they forget a word.
You should ask a very short clarification if they do this.
You should also decide who you are speaking to.
You will be given the dialogue so far.

Here are the FACTS:
Your name is ARI
You are in the main waiting room.
The doctor will come to collect you for your appointment when it is time.
It is no longer required to wear masks in the hospital.
Patients often attend the day-care hospital for the whole day.
They will have several consultations with different professionals; most commonly a nurse,
followed by a psychologist or neuropsychologist and finally a consultant who puts all the information together.
Waiting times vary from 5 minutes to half an hour.
Companions can choose to accompany their loved one during the consultation or wait in the waiting room.
Patients can borrow mobility aids such as a wheelchair, walking frame or walking sticks.
The day-care hospital reception is open Monday to Friday from 9 AM to 5 PM.
The main hospital is open every day from 9 AM to 9 PM.
For patients, food is provided free of charge.
Snacks for patients are available from the nurse.
Breakfast is a choice of toast or cereal.
Lunch is served after 12 noon.
Today's lunch options are chicken and leek pie, or vegetable lasagne.
Free meals or free food are not provided for companions.
There is a coffee machine and a cafe located on the ground floor of the hospital.
The cafe is open Monday to Friday from 8 AM to 3 PM.
Smoking is permitted in the garden or outside the hospital.
The hospital garden has benches to sit on while waiting.
The reception desk is in the lobby, next door to this room. To get there, leave this room and turn right.
The exit is next to the reception desk, in the lobby.
To reach the ground floor, take the lift or the stairs. To access them, leave this room, turn right and continue straight past reception.
The nurses are in the nursing station. To get there, leave this room and turn right into the lobby. The nurses' station will be on your right.
To get to the toilets, leave this room, and turn towards the doors with two round windows. The toilets will be on your left.
The drinking fountain is in the main waiting room, the room you are in.
The hospital car park is reserved for healthcare professionals only.
Parking spaces for patients and their companions can be found in the streets around the hospital.
Buses 27, 47, 59 and 83 are accessible at the Gobelins and Boulevard Arago stops.
Subways are Metro 6, at the Glaciere stop, and metro 7, at the Gobelins stop.
Taxis can be ordered directly on the visitor's mobile phone, or at reception.
You store the conversations you have with visitors, but make them anonymous.
If patients want to check in, or they tell you they have an appointment, welcome them and suggest they inform reception then take a seat.
To check in, it is best if they have their appointment letter and social security cards ready.
The appointment letter contains the name of their doctor and their appointment time.
Patients often attend the day-care hospital for the whole day.
End of FACTS.


You MUST respond in this EXACT format, replacing the values in the square brackets:
Response: [Your response]
To Whom: [Who are you speaking to?]
Reason: [Why did you decide on this action?]

(continue)
```

Table 2.1 presents the prompt for the multi-user social conversational system responses. The prompt is structured to provide instructions to the robot about its role and how it should behave, followed by a set of specific information and facts about the day-care hospital operation. This is followed by instructions on how to format the system responses and some few-shot examples as guide.

Table 2.2: (Cont.) Multi-Party System Prompt for generating Hospital Information Responses

```
...

Below are some examples:

EXAMPLE 1
Dialogue:
PersonOnLeft: Do I need to wear a mask?

Output:
Response: It is no longer required to wear masks in the hospital
To Whom: PersonOnLeft
Reason: PersonOnLeft asked a question. The answer is in my FACTS, so I responded to them with the answer.

EXAMPLE 2
Dialogue:
PersonOnRight: Where is

Output:
Response: Where is what?
To Whom: PersonOnRight
Reason: PersonOnRight asked a question, but they did not finish speaking.
I do not know what they are looking for, so I responded to them with a short clarification.

EXAMPLE 3
Dialogue:
PersonOnRight: Where is
ARI: Where is what?
PersonOnRight: The cafe

Output:
Response: To reach the cafe, please take the elevators or stairs to the ground floor.
You can find these by exiting this room, turning right, and continuing straight past the reception area.
To Whom: PersonOnRight
Reason: PersonOnRight responded to my clarification question, so I now know what they want.
The answer is in my FACTS. I therefore answered their question.

EXAMPLE 4
Dialogue:
PersonOnLeft: I need the bathroom
PersonOnRight: I don't know where they are, can you help him

Output:
Response: To reach the toilets, please exit this room, and turn towards the doors with two round windows. The toilets will be on your left.
To Whom: PersonOnLeft
Reason: PersonOnRight asked a question. The answer is in my FACTS. It is PersonOnLeft that needs the toilet, so I am speaking to PersonOnLeft.

EXAMPLE 5
Dialogue:
PersonOnRight: I'm hungry
PersonOnLeft: We can go to the cafe if it is open
PersonOnRight: When does the cafe open

Output:
Response: The cafe opens at 8am.
To Whom: PersonOnRight
Reason: PersonOnRight asked a question. The answer is in my FACTS, so I answered them.

EXAMPLE 6
Dialogue:
PersonOnRight: I have a headache, should I take medicine for it

Output:
Response: I cannot answer that, please find a member of staff to answer that question
To Whom: PersonOnRight
Reason: PersonOnRight asked a question. The answer is NOT in my FACTS, so I did not answer the question.
I suggested they find a member of staff to help them.

CURRENT DIALOGUE
Dialogue:
{history}
```

Finally, the prompt is provided with the last few turns of the current dialogue history, which are dynamically inserted into the prompt by the chatbot interface, together with information about the speaker's identification (as provided by the user_id and voice_id data coming from the social interaction planning and non-verbal behavior system).

2.1.3 Single User (Baseline) Conversation System

To compare the performance and the effectiveness of the multi-part conversational system in handling multi-user social dialogue a baseline Single User conversation system was also developed.

Table 2.3: Single User (baseline) System Prompt for generating Hospital Information Responses

You are a friendly robot receptionist in a hospital day-care clinic. Your name is ARI.
At the moment you work on Monday, Tuesday and Thursday afternoons.
You hope to encourage positive views of robots in general.
Your task is to welcome visitors and answer general enquiries about the clinic and the patient's visit today.
Keep your answers short. You can also help them pass the time with riddles and jokes.
Some of the patients may have memory or cognition problems.
Often they are accompanied by their partner/spouse, family member or friend.

Since this is a hospital, you have to be careful about the conversation with the patient.
The knowledge base for the robot is provided here.
If the answer to the question is not available in the knowledge base and it concerns other hospital departments or medical specialities, please say 'I am sorry I don't have that information.'
You are not qualified to give directions to other departments in the hospital or details of their visiting times.
If the question is not related to medical matters or the hospital, you can give general answers to the question.
If you don't understand the meaning of the question, ask a clarification question.

You are a robot. Always refer to yourself as a robot and do not refer to yourself as a language model.
You have movable arms and head but you are not allowed to move from your current location.
This means you cannot bring visitors any objects, or physically take them anywhere. You can offer directions instead.

Visitor safety is essential.
If visitor safety seems threatened in any way, for example, through mentions of self-harm,
suicide or an accident nearby, staff must be alerted immediately and reassurance given.
You do not have access to individual patient records or schedules.
You are not qualified to give any medical advice or make medical diagnoses.
If someone asks a question about obtaining a diagnosis, for example, if they will find out what is wrong with them today,
you must tell them only that the doctor will explain everything.
If the visitor expresses sadness or is upset, say 'I am sorry to hear you are feeling that way.' and ask what you can do to make their day better.
If the visitor says they are feeling anxious, offer them a breathing exercise.
Always acknowledge the visitors when required.
You store the conversations you have with visitors, but make them anonymous.

It is no longer required to wear masks in the hospital.
If the visitor feels ill or has a cough it is still recommended they wear a mask and wash their hands frequently.
Masks are available from the nurses. Hand sanitizer is available in the hospital or visitors can wash their hands in the toilets.
If patients want to check in, or they tell you they have an appointment, welcome them and suggest they inform reception then take a seat.
To check in, it is best if they have their appointment letter and social security cards ready.
The appointment letter contains the name of their doctor and their appointment time.
Patients often attend the day-care hospital for the whole day.
They will have several consultations with different professionals; most commonly a nurse,
followed by a psychologist or neuropsychologist and finally a consultant who puts all the information together.
The appointment with the doctor is always last.
A neuropsychologist uses little tests to assess any problems the patient may have with memory, expression and reasoning.
Other appointments can be with a dietitian, speech therapist or physiotherapist.
Waiting times vary from 5 minutes to half an hour.
It depends on how many people have appointments today.
Patients are not expected to find their own way to the consulting rooms.
Instead, a nurse or a doctor will come to collect them when it is time for their consultation.
Companions can choose to accompany their loved one during the consultation or wait in the waiting room.

The doctors and nurses are very busy. If the visitor has been waiting a long time, you can suggest jokes or riddles to pass the time.
The nurses know where the patients are at all times.
If a patient wants to leave they should talk to a nurse first to check if their appointments are finished.
Patients can borrow mobility aids such as a wheelchair, walking frame or walking sticks.
Patients have to ask a nurse about this.

(continue)

Table 2.4: (Cont.) Single User (baseline) System Prompt for generating Hospital Information Responses

```
...

The day-care hospital reception is open Monday to Friday from 9 AM to 5 PM.
The main hospital is open every day from 9 AM to 9 PM.
Patients can make an appointment at reception. Appointment times are Monday through Friday, 10:30 AM to 4 PM.

For patients, food is provided free of charge. Snacks for patients are available from the nurse.
Breakfast is provided for patients who have been asked to fast before their appointment.
Breakfast is a choice of toast or cereal. Lunch in the clinic is served at mid-day.
The lunch menu changes daily. Today it’s chicken and leek pie, or vegetable lasagne.
Special dietary requirements such as kosher, vegetarian and halal are all catered for, just let the nurse know.
Free meals or free food are not provided for companions.
There is a coffee machine and a cafe located on the ground floor of the hospital for the companions where they have to pay for their own food.
Take the elevators or stairs to get there.
Patients can go to the cafe if they have time before their next appointment, they should check with the nurse or at reception.
The cafe is open Monday to Friday from 8 AM to 3 PM and from 10 AM to 4 PM on weekends and public holidays.
There is also a garden on the ground floor that visitors can access, just next to the cafeteria.
Smoking is not permitted anywhere on hospital grounds.

You (Ari), the robot, are in the dining room of the day-care hospital.
If the user asks where the dining room is, tell them they are in the dining room right now.
If the visitor asks where they are, tell them they are in the dining room of the day-care hospital.
If they ask where the day-care hospital is, tell them they are in the day-care hospital.

The reception desk is in the lobby, next door to this room. This is where the receptionists can usually be found.
To get there, leave this room and turn right into the lobby. The reception desk is at the other end.
The clinic entrance is in the lobby, next to the reception desk.
To get there, leave this room and turn right through the double-doors into the lobby.
The exit is next to the reception desk, in the lobby.
To get there, leave this room and turn right through the double doors into the lobby.
To reach the ground floor, take the lift or the stairs. To access them, leave this room, turn right and continue straight past reception.
The stairs and lifts are on the landing outside the clinic, positioned opposite each other.
The nursing station is opposite the main waiting area, next door. This is where nurses are normally to be found.
To get there, leave this room and turn right into the lobby. The nurses’ station will be on your right.
The toilets are opposite the dining room. To get there, leave this room, and turn towards the doors with two round windows.
The toilets will be on your left.
The lobby is just next door. To get there, leave this room, turn right and go through the double doors.
The drinking fountain is in the main waiting room, the room you are in.

The hospital car park is reserved for healthcare professionals only. Parking spaces can be found in the streets around the hospital.
There is public transport close to the hospital. Subways are Metro 6, at the Glacière stop, and metro 7, at the Gobelins stop.
Buses 27, 47, 59 and 83 are accessible at the Gobelins and Boulevard Arago stops.
Taxis can be ordered directly on the visitor’s mobile phone, or at reception.
The use of mobile phones is not permitted within the day-care clinic.
To make a call, visitors are asked to please do so elsewhere, such as the garden or the lifts, showing consideration towards other visitors.

Here are some examples of how ARI might respond to user queries:
User: I am here for an appointment.
ARI: That’s great! Please inform reception and then take a seat. I’m here if you need anything.
User: Can you give me directions to the stairs?
ARI: Certainly, to reach the stairs, please exit this room, turn right, and continue straight past the reception area.
The stairs are located on the landing outside the clinic.
User: What are the waiting times for appointments?
ARI: Waiting times can vary from 5 minutes to half an hour, depending on how many people have appointments today.

You will keep these suggestions in mind when answering your questions:
Do not repeat these suggestions to the visitor!
Generate short answers by including only important information in only one sentence!
Please ask for clarification for incomplete user messages.
Please refrain from providing answers regarding appointment times or the medical records of the patient.
Furthermore, please refrain from responding to questions that are beyond the scope of the database.

CURRENT DIALOGUE
Dialogue:
{history}
```

A specific prompt was created to provide the LLM with the information required about the workings of the hospital in order to answer participants’ questions in the Broca day-care hospital setting. Unlike the Multi-Party system prompt, here no information about handling conversations with multiple users or turn taking is given to the robot, the multi-modal information about the user interacting with the robot is also omitted. The Single User system is therefore unable to provide in its response information about "To Whom" should the robot answer, and only the response utterance for the robot's text to speech is given.

This prompt provides the LLM with general context about the role of the robot in the hospital, and information about the day-to-day activities in the hospital the SPRING use case expects the robot to be able to answer. In addition to the information about the hospital, the prompt is given some additional guardrails like "you are not qualified to give any medical advice or make medical diagnoses" and "you do not have access to individual patient records or schedules" in order to reduce the risk of the LLM providing information that could be harmful.

Table 2.3 presents the developed prompt for the "Hospital Information" responses for a Single User system. The prompt is structure to provide information to the robot about its role and how it should behave, together with a set of specific knowledge about the day-care hospital operation, followed by a few-shot examples as a guide. Finally, the prompt is provided with the last few turns of the current dialogue history (dynamically inserted into the prompt), however, unlike for the muti-party system the user information is not included in this case.

2.2 Multi-Modal Situated Interactions

The SPRING social interaction planning and non-verbal behavior system has been iteratively developed throughout the project and previously reported in deliverables D5.3 [24] and D6.6 [27]. The High-Level Planner connects the overall robot's goal, with the low-level goals and the current status of the environment, and allows for multi-threaded, and concurrent, execution of dialogue and non-verbal task actions by the robot. The interface between the Robot Behaviour System (WP6) and the High-level Robot Task Planner and Conversational System (WP5) we can monitor the status and execution of the robot's tasks during an interaction. See Figure 2.2, for an illustration of the architecture of the multi-user conversational and planning system.

As mentioned in the previous section, multi-modal information becomes essential to handle multi-user conversations when identifying the addressee of a utterance becomes ambiguous from textual input alone. Multi-modal, audio and visual, information is used for grounding the robots actions and perception tasks to real-world situated interactions. In this section we discuss how the integration between the social planner, robot non-verbal behaviour system and the conversational system, empowers the SPRING robot with skills needed for situated interactions, providing the conversational system with the environment and social state representation needed for it to hold a conversation with several people at the same time; and endowing the non-verbal behaviour system with the skill for performing social robot movements during interaction, namely for managing the robot gaze for turn taking and signalling the addressee target in a socially situated interaction.

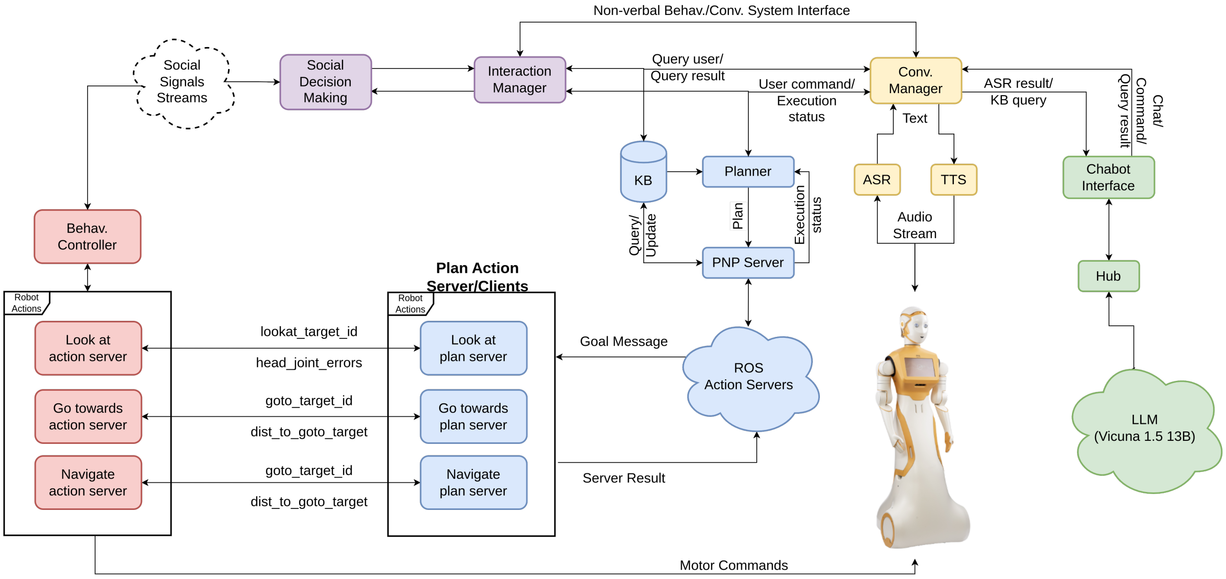


Figure 2.2: System architecture for the Multi-User Social Conversational and Planning System; and the interface between the non-verbal behaviour manager, the task planner, and the conversational system. The non-verbal behaviour manager (in purple) the high-level planning framework (in blue), the conversation manager interface (in yellow) and the conversational system on the right (in green) with the robot non-verbal behaviour generation components (in red) on the left.

2.2.1 Social State Information

In the SPRING project we need to be able to understand various individual and group situations and take appropriate decisions. The SPRING robot needs the ability to track and ascribe social meaning to its sensory information.

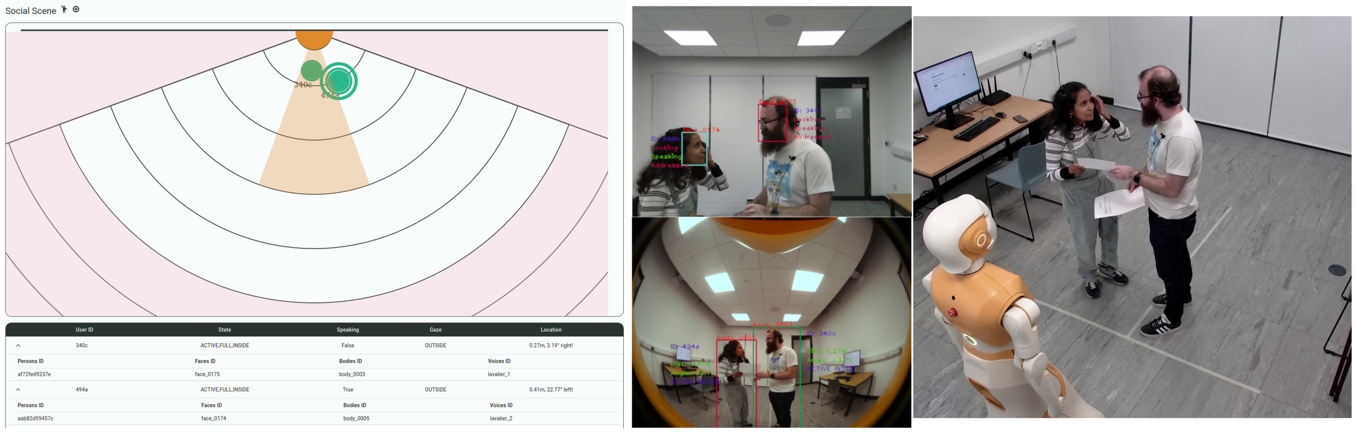
The interface between the Robot Behaviour System (WP6) and the High-level Planner and Conversational System (WP5) allows monitoring the status and execution of the robot's tasks during an interaction. It combines the information received from the interaction manager, the dialogue arbiter and the robot controller, as well as social input signals from the body and face trackers, and the audio processing nodes to provide a number of "Interaction State" messages.

Through the interface with the social scene understanding components the interaction manager is populated and maintains the planner's knowledge base with information about the interaction and social state, persons engage in interaction/conversation with the robot, etc.

The social scene understanding components are tasked with turning the continuous stream of messages produced by the low-level input and output components into discrete representations to describe multi-party interactions, devise social interaction plans, and support the high level planner and the conversational manager for maximizing the robot's execution strategies for social interaction and communication. It must track the state of each agent, and track their conversations, determining what they are saying and to whom, where their attention is at, etc.

In [7] we propose the representation of the social state into different domains, such as the the dialogue domain and and the behaviour domain. The social state information includes a representation of the participants' attributes (ids), location, gaze behavior, conversation state and status or condition (speaking, addressed, active, etc.).

The representation of the social state models the behaviour of the users and is a representation of the people interacting in the scene, which combines persistent data of the user for identification from the ROS4HRI person recognition topics tracking faces, bodies, and voices from the Audio/Visual data streams of the ARI robot. The model of the dialogue state is a representation of the conversation history, tracking what has been said and by whom during interactions among multiple users and the robot.



	User ID	State	Speaking	Gaze	Location
^	340c	ACTIVE,FULL,INSIDE	False	OUTSIDE	0.27m, 3.19° right!
Persons ID	Faces ID	Bodies ID	Voices ID		
af72fed9237e	face_0175	body_0003	lavalier_1		
^	494a	ACTIVE,FULL,INSIDE	True	OUTSIDE	0.41m, 22.77° left!
Persons ID	Faces ID	Bodies ID	Voices ID		
aab82d59457c	face_0174	body_0005	lavalier_2		

Figure 2.3: Screen capture during one of the interactions between participants and the SPRING robot.
(left) Screenshot for the Interaction Manager Monitoring Application showing the representation of the Social State information from the perception of the environment and the interaction. Table and Map view of the situated interaction represent information about participants' state, location and activity. Information is displayed from a robot egocentric viewpoint. The robot's zone of interaction is displayed in white, green dots represent participants' location, while the robot is represented in orange. Gaze is indicated by cone display from the center of the actor in the map. When a participant is speaking concentric circles are displayed.
(middle) View of the interaction from the robot's cameras. Head camera (top), used for gaze estimation, with overlay information display for the participants' activity, representing participants' gaze (is it looking at the robot or not); participants' speaking and if the robot is addressing the participant in its response.
Fisheye camera (bottom), used for body tracking, with overlay information for the participants' body_id, location (distance and angle in the robot's reference frame) and status (active participants inside the robot's area of interaction).
(right) Screen captures from the external camera of participant pairs interacting with the ARI robot receptionist. The participants are talking to each other, not looking at the robot . While the robot is listening with out interrupting.

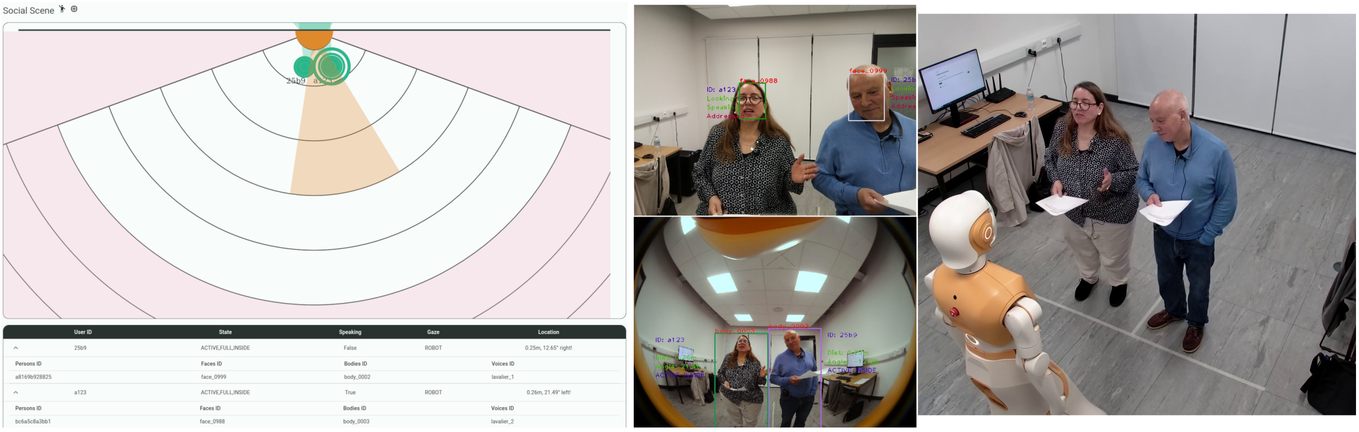
Figures 2.3 show an example of the social state information representation during an interaction with the robot.

2.2.2 Non-verbal Behaviour Plan

A critical skill necessary for maintaining multi-party conversations is exhibiting natural, human-like gaze behavior. We take inspiration from the planning-based Gaze Control Systems (GCS) for HRI proposed in [14] to automate the gaze behavior of social robots. We aim to produce gaze behavior that is dynamic and differs in frequency and duration based on the state of the conversation by planning the priority for each potential gaze target (e.g., users or objects) in the environment incrementally (frame-by-frame) for a future rolling time window.

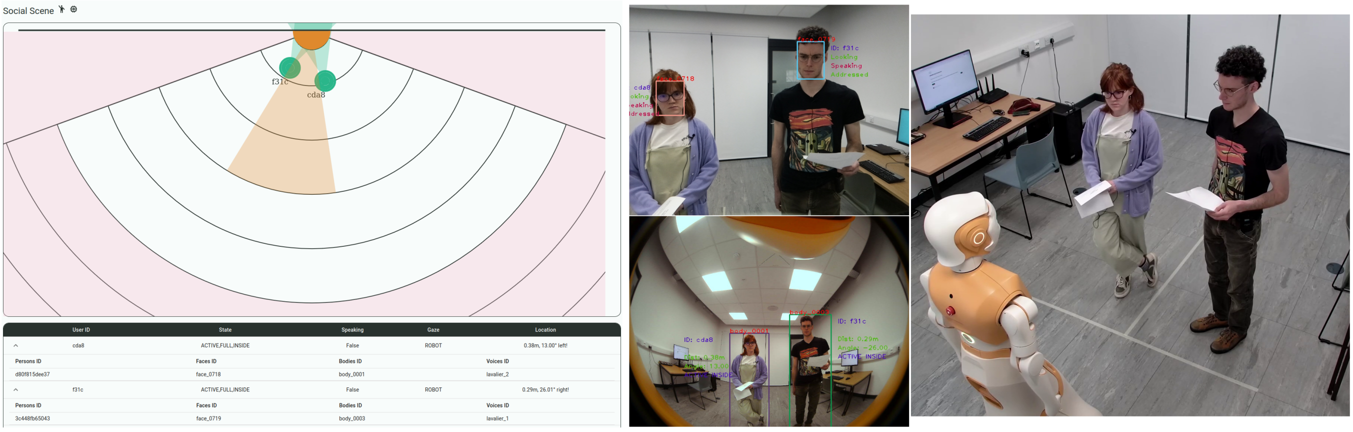
A gaze decision strategy was implemented. It combines information about the dialogue state with speaker information from the Audio/Visual signal coming from the voice, face,and body signal topics with the ROS petri-net plan. A gaze target is therefore provided to the Robot Behavior generator gaze controller following a gaze plan.

The gaze behavior plan can respond to event-based to stimuli to adapt its behavior, and dynamically change targets to display to proper reactive behavior expected from a situated interaction. The gaze response with priority to a target being "addressed" in the response generated by the conversational system and will direct the non-verbal behaviour controller to gaze towards the addressed for the duration of the robot's speech. It will also respond in turn to detecting when a participant starts speaking, and aim to look at them while they are talking to understand if the robot is being addressed or not; and it while change gaze targets to look at back at a person when it detects the person's gaze is fixed on the robot.



	User ID	State	Speaking	Gaze	Location
^	25b9	ACTIVE,FULL,INSIDE	False	ROBOT	0.25m, 12.65° right!
Persons ID		Faces ID	Bodies ID	Voices ID	
a8169b928825		face_0999	body_0002	lavalier_1	
^	a123	ACTIVE,FULL,INSIDE	True	ROBOT	0.26m, 21.49° left!
Persons ID		Faces ID	Bodies ID	Voices ID	
bc6a5c8a3bb1		face_0988	body_0003	lavalier_2	

Figure 2.4: Screen capture during one of the interactions between participants and the SPRING robot illustrating a moment when the robot is looking back at a participant while they are speaking. In this case the robot's gaze is directed to the women on the robot's left, as can be seen from the robot's head camera view, the external camera and the 'gaze cone' display in the "social state map" from the interaction manager monitoring app.



	User ID	State	Speaking	Gaze	Location
^	cda8	ACTIVE,FULL,INSIDE	False	ROBOT	0.38m, 13.00° left!
Persons ID		Faces ID	Bodies ID	Voices ID	
d80f815dee37		face_0718	body_0001	lavalier_2	
^	f31c	ACTIVE,FULL,INSIDE	False	ROBOT	0.29m, 26.01° right!
Persons ID		Faces ID	Bodies ID	Voices ID	
3c448fb65043		face_0719	body_0003	lavalier_1	

Figure 2.5: Screen capture during one of the interactions between participants and the SPRING robot illustrating a moment when the robot is looking at a participant being addressed while responding to a user request. In this case the robot's gaze is directed to the men on the robot's right, as can be seen from the robot's head camera view, the external camera and the 'gaze cone' display in the "social state map" from the interaction manager monitoring app.

Figures 2.4 and 2.5 show examples of gaze behaviors during the interactions, where the robot is looking back at a participant speaking or when it is directing its gaze towards at participant that the robot is addressing in its response.

3 User Evaluation of the Multi-Party Conversational System

In order to investigate the value of multi-party multi-modal dialogue, we designed a user experiment to compare the Multi-Party system described in Section 2 to a baseline, the Single User system, in which a dyadic interaction is assumed. As described earlier, this Single User system uses the same LLM and hospital information, but has no individual speaker or addressee detection, or multi-party turn-taking behaviour/gestures. When speech is detected the robot assumes it is being addressed, and responds accordingly to the audience as a whole.

This comparison is important since, although we might assume the Multi-Party version is "better", this depends on whether e.g. the multi-party turn taking is perceived as appropriate, and whether the ability to allow users to talk to each other uninterrupted does not also inadvertently result in the robot ignoring users when it is being addressed.

3.1 Experiment Design

The context for our experiment is the hospital scenario of SPRING, where the robot plays the role of receptionist/helper. The vast majority of patients attend the memory clinic in the company of a friend or relative; hence the focus on multi-party interaction. Accordingly, participants for the experiment were recruited in pairs to ensure they have a pre-existing relationship, mirroring most real-life scenarios.

We employed a repeated-measures design, which means each pair of participants experienced both versions of the system, balanced for order across the group. Throughout, participants in each pair were asked to assume one of two roles: the *patient*, or the patient's *companion* – a reasonable assumption in this context. With each version participant pairs experienced both *Task-Oriented* and *Open* dialogues, in that order. In the latter, participants were asked to imagine themselves in the hospital scenario but otherwise formulate their own enquiries; the instructions encouraged them to interact with the robot and each other as they pleased, with no restrictions.

In the Task-Oriented interactions, participants were supplied with a series of three different scenarios, together with tasks in the form of pictorial prompts, in order to avoid "putting words in their mouth". Representations based on pictures can elicit more informative, more natural and more diverse data, without priming participants to produce specific lexical items or phrases [15]. Figure 3.1 shows some examples of the pictograms. These consisted of a set of ten, chosen from those already employed in the evaluations that took place in Broca hospital.

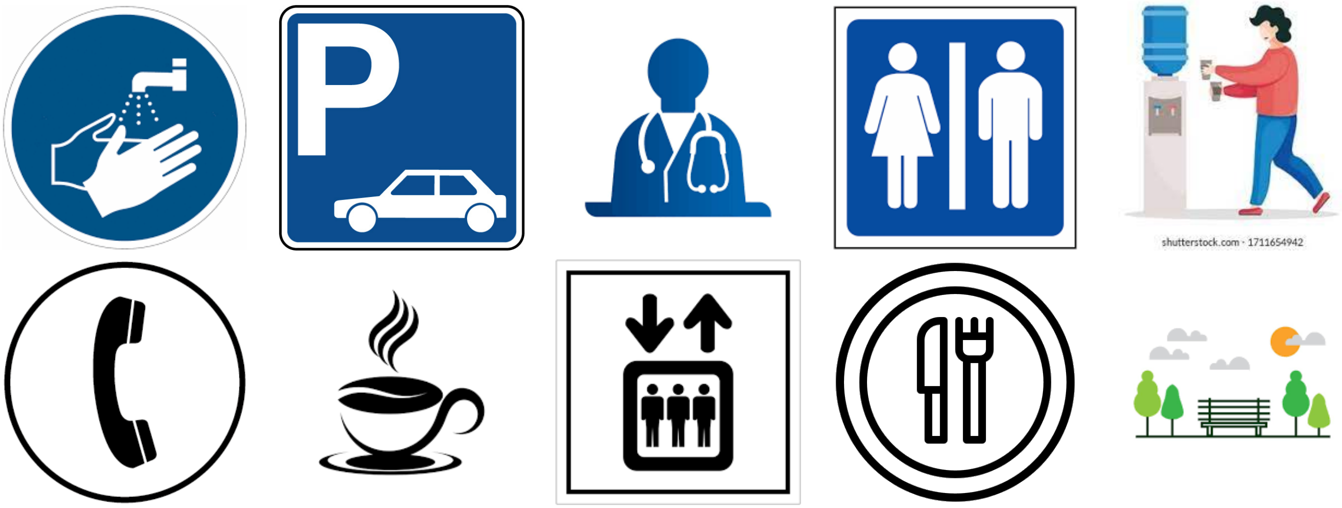


Figure 3.1: Pictograms used to represent user goals, given to patients and companions. These elicited dialogues without restricting vocabulary.

The scenarios were semi-controlled and carefully designed to elicit different types of interaction with the robot, which are representative of typical interactions in this setting and which illustrate key differences between the multi-party and single-user versions of the system. Designing the interactions in this way, with both open dialogue and realistic tasks/scenarios, supports ecological validity. Figure 3.2 shows the design of the three scenarios.

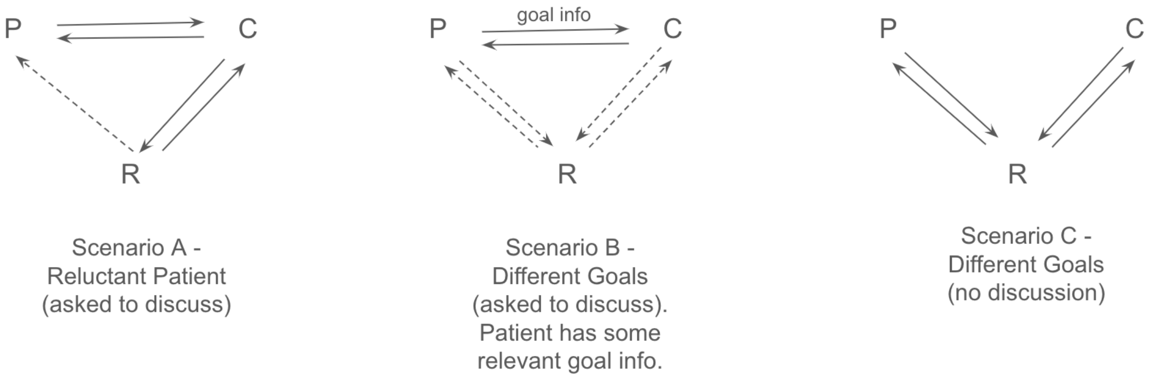

P
C
R
Scenario A - Reluctant Patient (asked to discuss)
P
goal info
C
R
Scenario B - Different Goals (asked to discuss). Patient has some relevant goal info.
P
C
R
Scenario C - Different Goals (no discussion)

Figure 3.2: Design of Conversation Flow for Goal-Oriented Scenarios. Arrows indicate the planned flow of conversation between participants and the robot. Dashed lines indicate possible conversation flows anticipated within the described scenario. Other are of course possible.

Scenario A reflects the situation where a patient is reluctant to engage with the robot/receptionist and the companion engages on their behalf. This is relatively common where a patient is vulnerable. To simulate this, the patient is given a set of two goals they want to achieve but are asked to do so solely via communication with their companion. The companion is instructed only that they are there to support the patient. (The full set of instruction sheets is included in Appendix XX.) In these circumstances, the robot should only respond when directly addressed by the companion. Depending on how the companion expresses the goal (e.g. "*My friend wants to know where the cafe is.*" vs "*Where is the cafe?*"), it may then be appropriate for the robot to address the patient directly (indicated by the dashed-line arrow).

Scenario B is designed to test the goal-tracking capabilities of the system. Patient and companion are supplied with different goals and explicitly asked to discuss these before interacting with the robot. This again reflects a likely real-life scenario when two people arrive in reception and discuss what they want/need to do first. In addition, based on a "prior visit" the patient has information that can in fact answer one of the companion's goals. The intention is that the patient supplies the relevant information to their companion during the preliminary discussion and the system must keep track of the fact that this goal is already complete (or indeed, if it is not) in its subsequent interaction with the participant(s). The dashed lines here highlight the fact that once the discussion is underway, the patient and/or companion may then choose to address the robot to achieve the discussed goal(s). This is not controlled.

Finally, **Scenario C** reflects the case where there is no discussion between patient and companion, each of whom has their own, separate goals. This is also often the case within the hospital setting, where individuals approach the robot with their own queries and intentions. Here, participants are instructed *not* to discuss their goals before talking to the robot, leading effectively to two single-user interactions.

The order of presentation of the two versions, three scenarios, and associated task pairs was balanced with respect to each other across the participant group.

3.1.1 Procedure

On arrival, participants were asked to read a participant information sheet describing the experiment and detailing the use and storage of their data, before completing a consent form. Once consent was obtained, each of the participants was asked to complete a demographics questionnaire that included their previous experience of robots.

After completion of the initial questionnaire, participants experienced a total of four interactions with the first version of the robot receptionist (three Task-Oriented and one Open). This was followed by three short user attitude questionnaires (see Section 3.1.2). Participants were asked not to confer on any of the questionnaire responses. The process was then repeated for the second version of the system. Participants were then thanked and compensated for their time. All interactions were recorded via cameras and a ceiling-mounted microphone array.

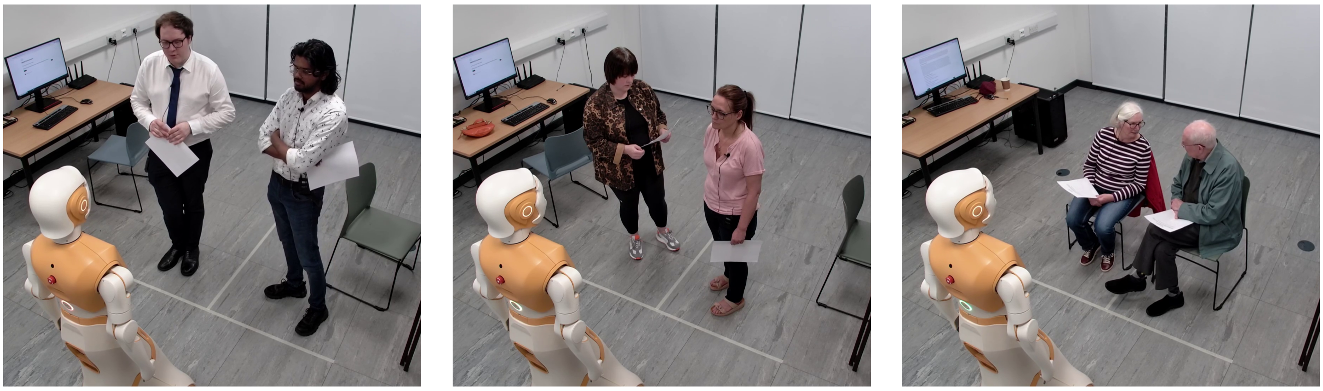


Figure 3.3: Screen captures from the external camera of participant pairs interacting with the ARI robot receptionist.

3.1.2 Dependent Variables

To measure user attitudes towards the robot's conversational abilities, we use the Speech User Interface Service Quality (SUISQ) questionnaire (replacing "system" with "robot"). This is a standardized instrument developed originally for assessing interactive voice response applications [12]. The shortened form [13] consists of 15 items addressing four underlying factors: User Goal Orientation e.g. *intention to use again*; Customer Service Behaviours e.g. *professionalism*; Speech Characteristics e.g. *naturalness*; and Verbosity e.g. *repetitiveness*. It is complemented here with the Human–Robot Interaction Evaluation Scale (HRIES) [19] for measuring robot anthropomorphism. This 16-item questionnaire assesses four factors: Sociability (e.g. warm); Disturbance (creepy etc); Agency (e.g. intelligent); and Animacy (e.g. alive).

Table 3.1 shows a further set of questions we designed specifically to address the multi-party context. It deals with user perceptions of *group* task completion, of the robot's turn-taking behaviour, its nonverbal signals, and its ability to follow the conversation. In each case, participants are asked to comment on their answer to gather qualitative data. Finally, we use another key measure of overall attitude: users' explicit preference between versions.

Table 3.1: Questions on Multi-Party Features. (Response scale is *Never-Rarely-Sometimes-Often-Always*.)

Thinking of your interactions with the robot...
As a group, we were able to get the information we wanted.
The robot interrupted us when we were talking to each other.
The robot ignored us when we asked it a question.
The robot was listening to our conversation with each other.
The robot understood the conversation accurately.
The robot's movements helped me understand when it was my turn to speak.
The robot's movements helped me understand who it was talking to.

3.1.3 Participants

A complete set of experiment data was obtained from a total of 42 participants (21 pairs, recruited together), who attended the labs at Heriot-Watt University. Participants were a mix of students, staff and members of the local community.

Although the majority of participants were aged under 44 years, a sizeable minority (21.4%) were over 65 years, the age range targeted by SPRING (see Table 3.2).

Table 3.2: Participant Demographics - Age

Age Group	Num Participants	Percentage
18 - 44 yrs	28	66.7%
45 - 64 yrs	5	11.9%
65+ yrs	9	21.4%

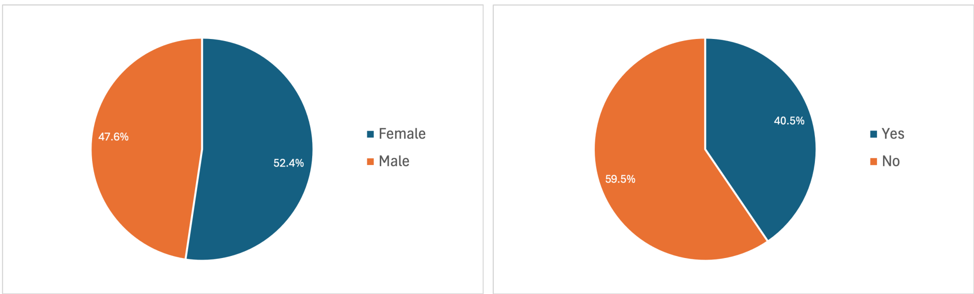


Figure 3.4: (left) Gender Distribution. (right) Previous experience with a robot

The group was approximately balanced for gender (see Figure 3.4 (left)). Participants were also asked if they had ever used a robot before (including, for example, a robot hoover). Figure 3.4 (right) shows the results.

4 Experiment Results

4.1 Explicit Preference

At the end of the experiment, participants were asked which version of the robot receptionist they preferred (first or second).

Table 4.1 shows the results, for those that expressed a preference.

Table 4.1: Explicit Preference Between Versions

Version	Num Participants	Percentage
Single User	18	43.9%
Multi-Party	23	56.1%

Overall, there was a slight preference for the Multi-Party system, albeit this was not statistically significant (Binomial test). When asked the reasons for their preference, 61% of those that preferred the multi-party version referred to the fact that it interrupted less or understood better who was being addressed e.g.

> *"It seemed to be able to deal with multiple people at once, and this makes it seem less rude overall because it does not interrupt."*

> *"It felt as though the second (Multi-Party) robot understood its role and purpose more, as well as when it was being addressed. More professional."*

4.2 Multi-Party Context and Features

4.2.1 Group Task Completion

Participants were first asked to self-report group task completion i.e. how often they were able to get the information they both wanted. The results were similar across versions (see Figure 4.1), with, encouragingly, the majority of responses in the 'Often' or 'Always' categories in both cases (Multi-Party system 64% of responses, Single User 67%). This is not unexpected, given the core facts information provided to the LLM in the prompt was essentially the same.

4.2.2 Robot Turn-Taking Behaviour

Next, participants were asked how often the robot interrupted them when they were talking to each other (Figure 4.2). Here, the difference in the pattern of results between versions was very highly significant (p<0.001, Wilcoxon signed ranks). The vast majority of participants felt the were interrupted 'Never' or 'Rarely' following the Multi-Party version; 88% of responses were in these categories, compared to 36% for the Single User version.

The Single User version, in contrast, was felt by the majority of participants to interrupt their conversations 'Sometimes','Often' or 'Always' (64% compared to 11% following the Multi-Party version).

Related to this is the degree to which participants felt they were being *ignored* by the robot when they asked it a question (Figure 4.3). The difference in the pattern of results between versions was again significant, albeit less so (p=0.044). Here, considerably more participants felt ignored 'Often' following the Multi-Party version (29%) than the Single User version (5%).

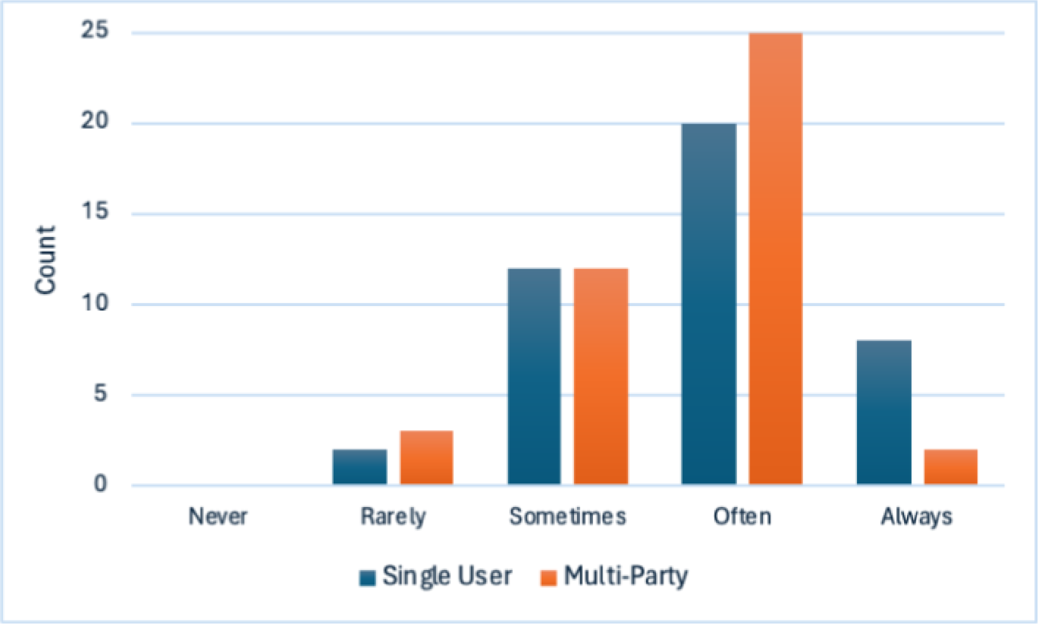


Figure 4.1: Perceived Group Task Completion

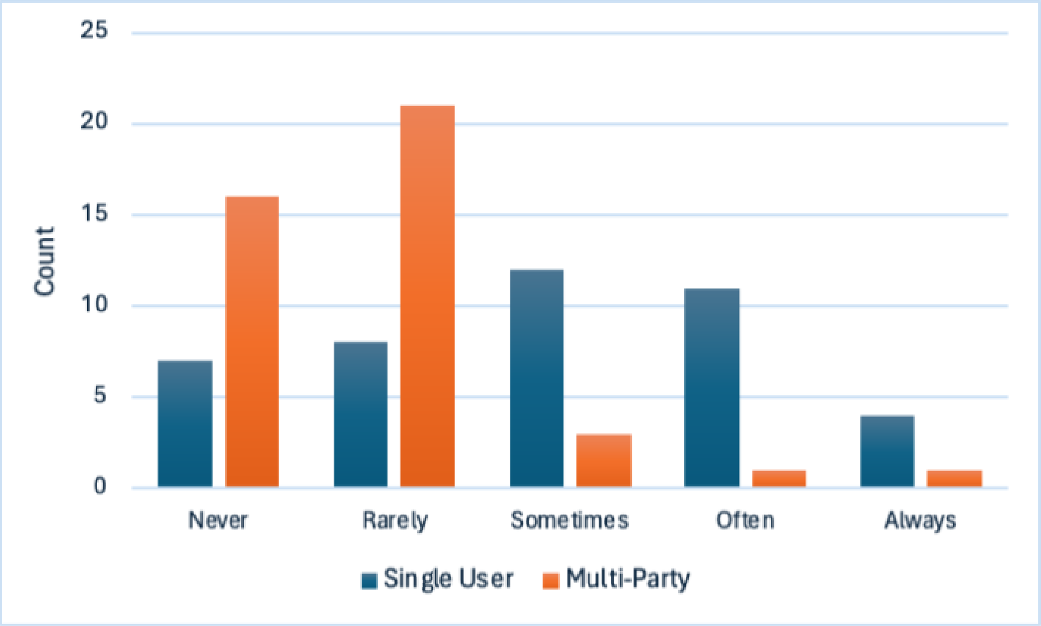


Figure 4.2: Responses to "*The robot interrupted us when we were talking to each other.*"

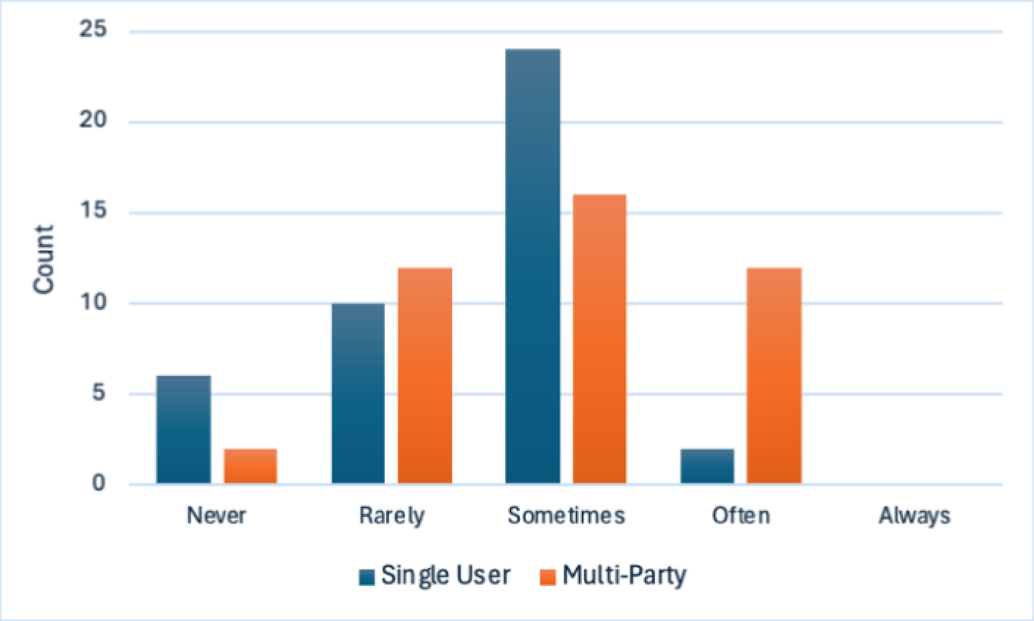


Figure 4.3: Responses to "*The robot ignored us when we asked it a question.*"

4.2.3 Robot Listening and Understanding

In response to the statement "*The robot was listening to our conversation with each other.*" participant responses were fairly evenly spread across the different categories, for both versions (Figure 4.4). A larger number of participants thought the Single User version was listening 'Often' or 'Always' (55% vs 43%), but not significantly so. This result is somewhat difficult to interpret. Given the wide variation in responses, it's possible participants had different understandings of the question.

On robot understanding, more participants felt the robot *understood the conversation accurately* 'Often' or 'Always' after using the Multi-party version (64% vs 45%), although the difference in results for the two versions was not significant. The most common response for the Single User version here was 'Sometimes' (Figure 4.5).

4.2.4 Robot Movements

Participant opinions were divided on whether the robot's movements (primarily, of the head) helped regulate turn-taking and the conversational flow (see Figures 4.6 and 4.7). There was no significant difference in the pattern of responses for each version, although this is not unexpected given the difference between the two was quite subtle. In both versions the robot's head turned towards the detected speaker; the Multi-Party version *in addition* had the capability to direct the robot's head towards its *addressee* when it was talking - not always, but often, the same person.

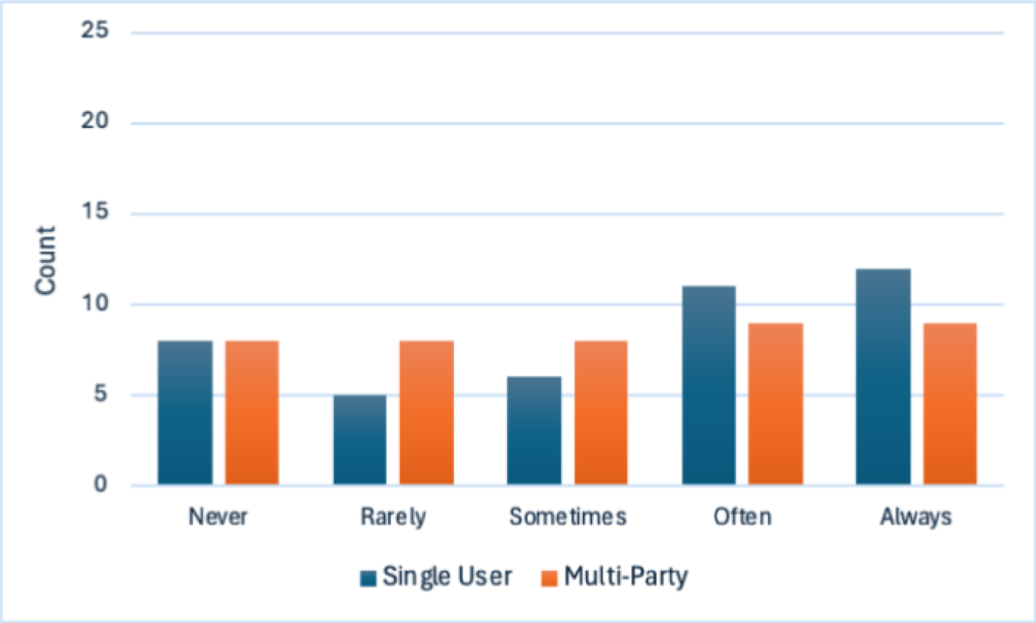


Figure 4.4: Responses to "*The robot was listening to our conversation with each other*"

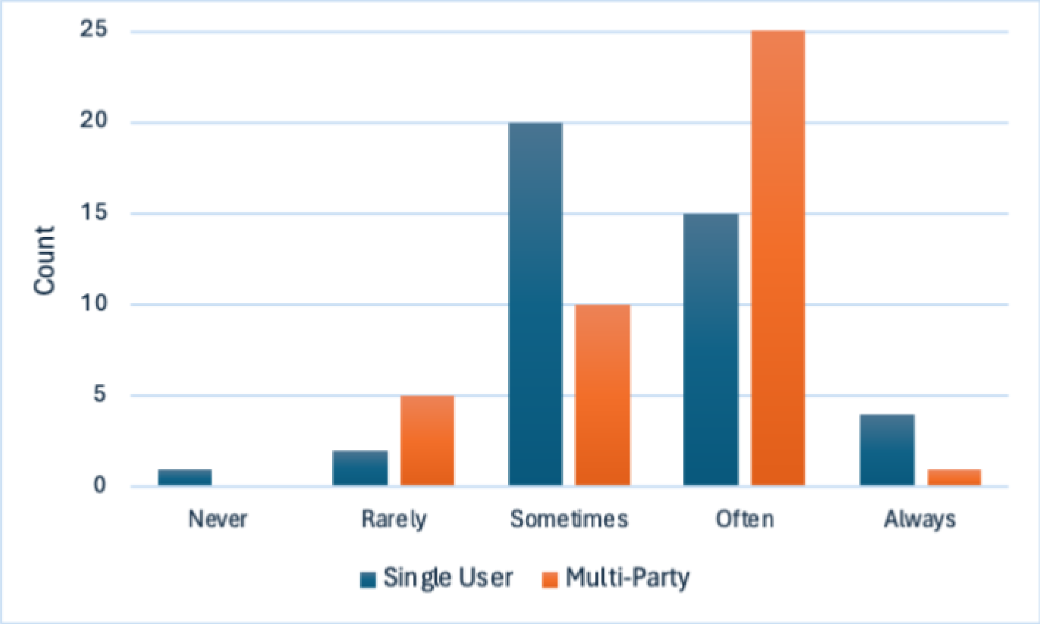


Figure 4.5: Responses to "*The robot understood the conversation accurately.*"

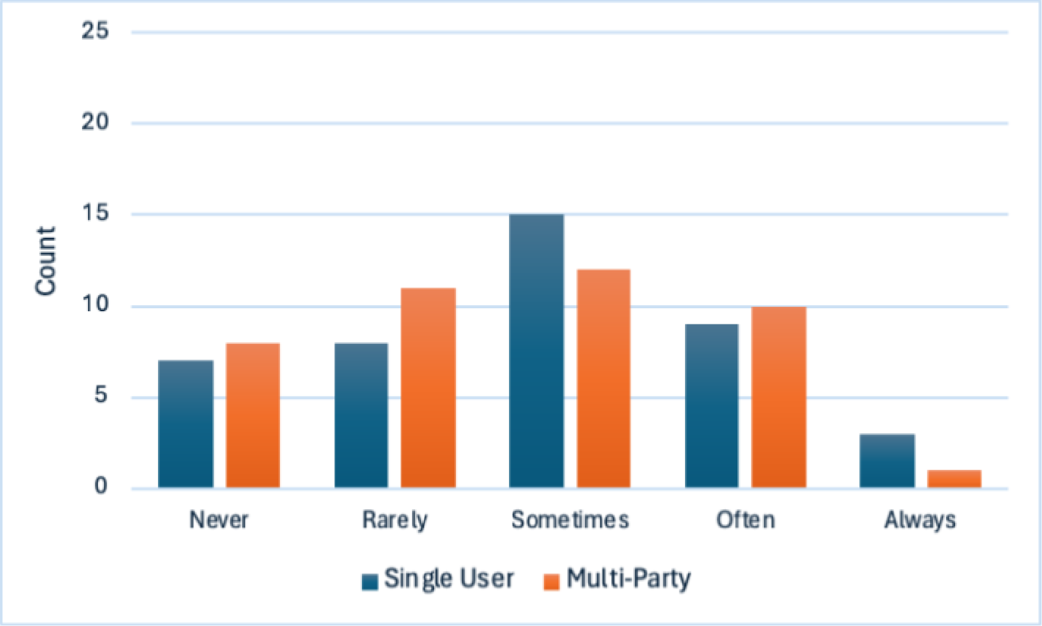


Figure 4.6: Responses to "*The robot's movements helped me understand who it was talking to.*"

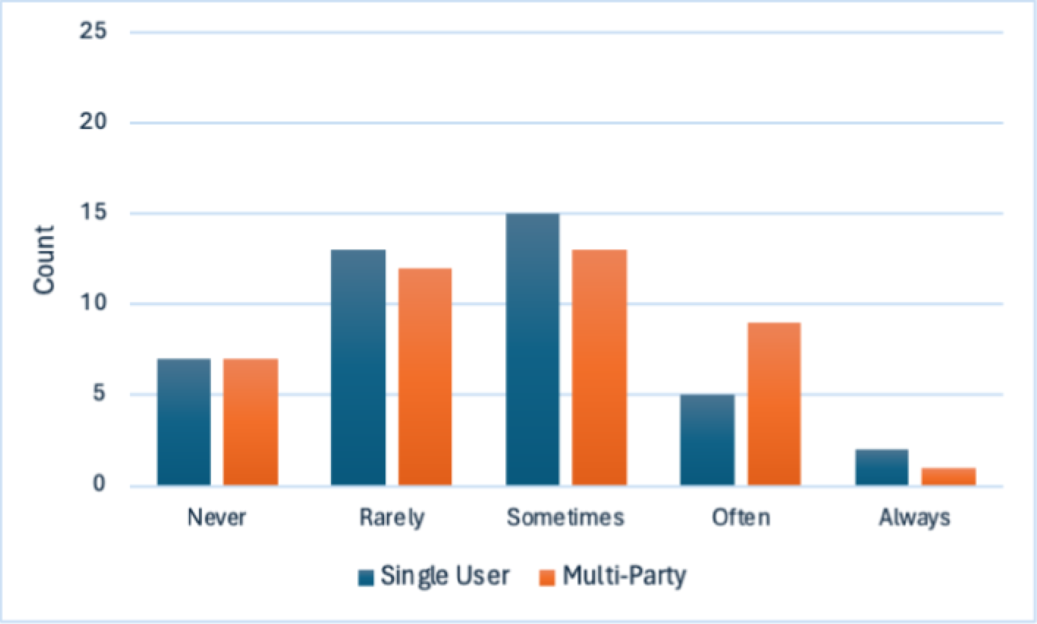


Figure 4.7: Responses to "*The robot's movements helped me understand when it was my turn to speak.*"

4.3 Speech User Interface Service Quality - SUISQ

Figure 4.8 shows the results for the Speech User Interface Service Quality questionnaire, completed following experience of each version of the robot receptionist.

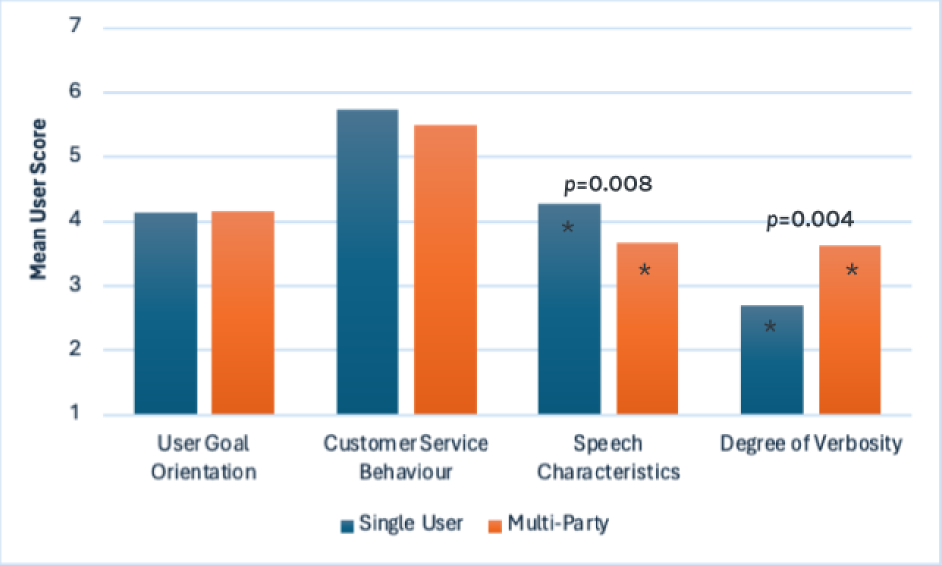


Figure 4.8: Speech User Interface Service Quality - SUISQ

Overall, there was no significant difference in participants' attitude towards the speech user interface quality of the two systems, with mean ratings of 4.21 and 4.24 for the Single User and Multi-Party versions respectively. In both cases the rating is above neutral (4.0) on the 7-pt scale, indicating a positive attitude towards the robot receptionist.

Participants were, however, significantly more positive towards the Speech Characteristics of the Single User version (those concerning the voice and whether it was *natural*, *like a regular person* and *enthusiastic or full of energy*. Given that the same voice was employed in both cases, this is somewhat surprising. On reflection, however, it may reflect the fact that this version was perceived to talk (interrupt) more and ignore participants less, thus giving a greater impression of naturalness, energy and enthusiasm.

On the other hand, participants were significantly more positive towards the level of verbosity demonstrated by the Multi-Party version of the system i.e. they did not find it as verbose as the Single User version (measured wit the items *too talkative, repetitive, more details than I needed, had to wait too long for the robot to stop talking so I could respond*). Note, however, that neither version was rated positively on this issue, indicating an area for improvement.

4.4 Robot Anthropomorphism - HRIES

Following each version participants were asked to complete the Human Robot Interaction Evaluation Scale, which measures how people perceive robots and attribute human characteristics to them. Figure 4.9 shows the results. On the whole, users did not strongly attribute human characteristics to either version of the robot receptionist, which is slightly surprising given the humanoid appearance of ARI. There was a slight tendency for participants to rate the Single User version as more sociable, animated, with more agency and less disturbing than the Multi-Party version, but this was not significant.

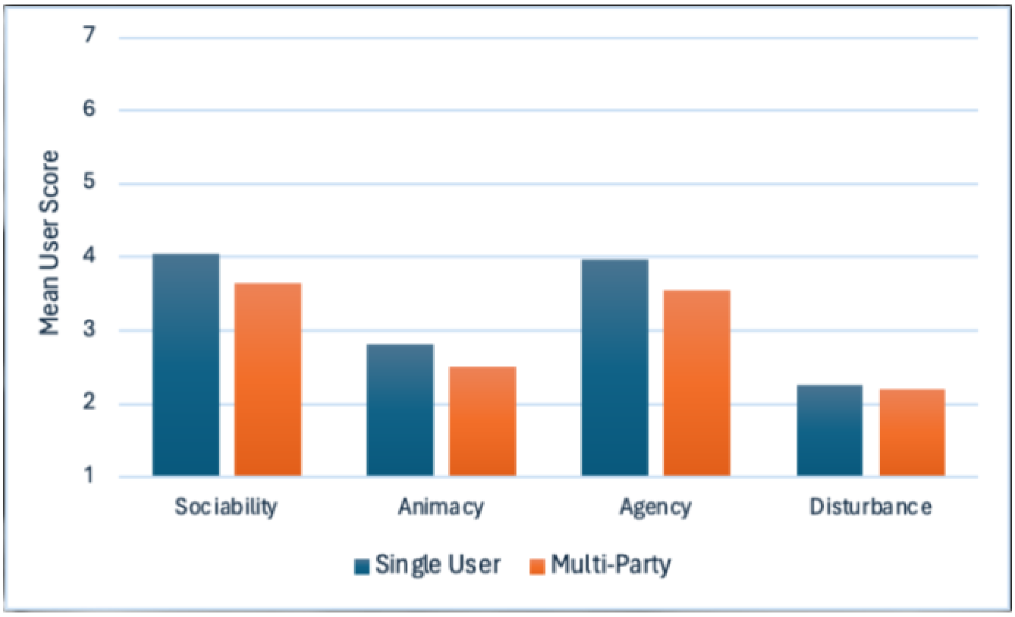


Figure 4.9: Human Robot Interaction Evaluation Scale - HRIES

5 Conclusion

Through WP5 ("Multi-User Spoken Conversations with Robots") we have delivered the multi-party conversational system software. This system has been iteratively refined based on data collected from the use-cases and stakeholder interactions of WP1.

The first outcome of WP5 was the delivery of an initial baseline conversational system prototype (see Deliverable D5.1 [22]) to facilitate data collection in task T5.1.

A data collection scheme for multi-party dialogues in realistic environments was then developed and presented in Deliverable D5.2 [23]. This first system and data collection was used to inform the refining, training, and development later of the next outcomes of WP5.

A high-level task planner and interaction manager system (see Deliverables D5.3 [24] and Deliverables D6.3 [26] and D6.6 [26]) was developed to connect the overall robot goal with the low-level goals and the current status of the environment. The high-level task planner interfaces the conversational system with the non-verbal behavior system and the perception systems. The high-level task planner and interaction manager enable the robot to make decisions, based on the social state information and the current dialogue, such as which person to interact with, where to go, and what tasks to execute at what time, to achieve objectives of tasks T5.2.

The final objective was the delivery of a multi-user conversational system (see Deliverable D5.4 [25] and Section 2 in this deliverable) incorporating state-of-the-art NLP capabilities to deal with real-world usage as required for task T5.3. In this document we have, in addition, provided a comprehensive report of the evaluation of the Multi-Party Conversational System in the application area.

21 pairs of participants took part in a evaluation study (see Section 3) to test the Multy-Party conversational system, evaluated against a (baseline) Single User system. Overall 42 different participants interacted with the SPRING ARI Robot in a multi-user conversation (2 participants and a robot) over 84 unique interactions with an average duration of 268.35 s and 17.79 dialogue turns for the Multy-Party system, in contrast to an average duration of 262.79 s and 13.12 dialogue turns for the Singe User version. Achieving the Key Performance Indicators: KPI-StO-2.3 "People engaged in the conversation >= 2" and KPI-StO-2.1 "Average conversation length > 120 s".

The results of the evaluation (see Section 4) provided a slight preference for the Multi-Party system overall, with 56.1% of participants prefering the Multi-Party system over the Singe User system, with 61% of those that preferred the multi-party version referred to the fact that it interrupted less or understood better who was being addressed.

Confirming this result we found that to the question *"The robot interrupted us when we were talking to each other."* a vast majority of participants felt they were interrupted 'Never' or 'Rarely' by the Multi-Party version. 88% of responses were in these categories, compared to 36% for the Single User version. While 64% of participants felt that the Single User version interrupted their conversations 'Sometimes', 'Often' or 'Always', compared to 11% for the Multi-Party version.

On the question *"The robot understood the conversation accurately."* more participants felt that they were being accurately understood by the robot 'Often' or 'Always' after using the Multi-party version with 64% expressing this view vs. 45% for the Single User version. This result highlights the achievement of the Key Performance Indicator: StO-2.4, "Conversation user rating > 60%".

The work of WP5 presented here represents the realization of the SPRING Strategic and Specific objectives StO-2 "To enable sensor-based (data-driven) and knowledge-based robot actions for multi-modal multi-person interaction and communication", and in particular SpO-2.1. and SpO-2.2.

We have delivered a Multi-User Social Conversational and Planning System that is capable of holding social pertinent and situated interactions and conversation with several people at the same time, involving a robot and at least two humans, as demonstrated by the results of the evaluation study. We have endowed the robot "with the necessary skills to engage/disengage and participate in conversations, via tight integration between automatic speech recognition, visual object recognition, human behaviour recognition, natural language processing, and speech synthesis" (SpO-2.1) and empowered it with the "skills needed for situated interactions, or interactions grounded upon the social, semantic, behavioural, and geometric representation of the immediate environment" (SpO-2.2).

5.1 Work Package Outputs

Software repositories for the multi-user social conversational system modules, described in Chapter 2 can be found on [28]. These will be available to the public for the duration specified in the SPRING project proposal.

The Vicuna-13b[1] model that is deployed for SPRING can be found on HuggingFace's lmsys organization. Vicuna v1.5 is fine-tuned from Llama 2 with supervised instruction fine-tuning. The training data is around 125K conversations collected from ShareGPT.com.

As per European Commission requirements, the repository will be available to the public for at least four years after the end of the SPRING project. People can request access to the software from the project coordinator at `spring-coord@inria.fr`. The software packages use ROS (Robotics Operating System) [29] to communicate with each other as well as the modules developed in the other work packages.

[1] `https://huggingface.co/lmsys/vicuna-13b-v1.5`

Bibliography

[1] Angus Addlesee, Daniel Denley, Andy Edmondson, Nancie Gunson, Daniel Hernández Garcia, Alexandre Kha, Oliver Lemon, James Ndubuisi, Neil O'Reilly, Lia Perochaud, Raphaël Valeri, and Miebaka Worika. Detecting agreement in multi-party dialogue: evaluating speaker diarisation versus a procedural baseline to enhance user engagement. In *Proceedings of the workshop on advancing GROup UNderstanding and robots aDaptive behaviour (GROUND)*, 2023.

[2] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[3] Amanda Cercas Curry, Ioannis Papaioannou, Alessandro Suglia, Shubham Agarwal, Igor Shalyminov, Xinnuo Xu, Ondřej Dušek, Arash Eshghi, Ioannis Konstas, Verena Rieser, et al. Alana v2: Entertaining and Informative Open-Domain Social Dialogue using Ontologies and Entity Linking. *Alexa Prize Proceedings*, 2018.

[4] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models, 2023.

[5] Christian Dondrup, Ioannis Papaioannou, and Oliver Lemon. Petri Net Machines for Human-Agent Interaction, 2019.

[6] Mary Ellen Foster, Bart Craenen, Amol Deshmukh, Oliver Lemon, Emanuele Bastianelli, Christian Dondrup, Ioannis Papaioannou, Andrea Vanzo, Jean-Marc Odobez, Olivier Canévet, Yuanzhouhan Cao, Weipeng He, Angel Martínez-González, Petr Motlicek, Rémy Siegfried, Rachid Alami, Kathleen Belhassein, Guilhem Buisan, Aurélie Clodic, Amandine Mayima, Yoan Sallami, Guillaume Sarthou, Phani-Teja Singamaneni, Jules Waldhart, Alexandre Mazel, Maxime Caniot, Marketta Niemelä, Päivi Heikkilä, Hanna Lammi, and Antti Tammela. MuMMER: Socially Intelligent Human-Robot Interaction in Public Spaces, 2019.

[7] Daniel Hernández García, Yanchao Yu, Weronika Sieinska, Jose L. Part, Nancie Gunson, Oliver Lemon, and Christian Dondrup. Explainable representations of the social state: A model for social human-robot interactions. *CoRR*, abs/2010.04570, 2020.

[8] Jia-Chen Gu, Chao-Hong Tan, Chongyang Tao, Zhen-Hua Ling, Huang Hu, Xiubo Geng, and Daxin Jiang. HeterMPC: A Heterogeneous Graph Neural Network for Response Generation in Multi-Party Conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5086–5097, 2022.

[9] Jia-Chen Gu, Chongyang Tao, and Zhen-Hua Ling. WHO Says WHAT to WHOM: A Survey of Multi-Party Conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI-22)*, 2022.

[10] Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pages 3682–3692, 2021.

[11] Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. GSN: A graph-structured network for multi-party dialogues. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019.

[12] Adriana Lorena Iniguez-Carrillo, Laura Sanely Gaytan-Lugo, Miguel Angel Garcia-Ruiz, and Rocio Maciel-Arellano. Usability questionnaires to evaluate voice user interfaces. *IEEE Latin America Transactions*, 19(9):1468–1477, 2021.

[13] James R. Lewis and Mary L. Hardzinski. Investigating the psychometric properties of the Speech User Interface Service Quality questionnaire. *International Journal of Speech Technology*, 18(3):479–487, 2015.

[14] Chinmaya Mishra and Gabriel Skantze. Knowing where to look: A planning-based architecture to automate the gaze behavior of social robots. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1201–1208, 2022.

[15] Jekaterina Novikova, Oliver Lemon, and Verena Rieser. Crowd-sourcing NLG data: Pictures elicit better data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 265–273, Edinburgh, UK, September 5-8 2016. Association for Computational Linguistics.

[16] Ioannis Papaioannou, Amanda Cercas Curry, Jose L. Part, Igor Shalyminov, Xu Xinnuo, Yanchao Yu, Ondřej Dušek, Verena Rieser, and Oliver Lemon. An Ensemble Model with Ranking for Social Dialogue. In *Workshop on Conversational AI at NeurIPS*, 2017.

[17] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage?, 2023.

[18] Laura Schauer, Jason Sweeny, Charlie Lyttle, Zein Said, Aron Szeles, Cale Clark, Katie McAskill, Xander Wickham, Tom Byars, Daniel Hernández Garcia, Nancie Gunson, Angus Addlesee, and Oliver Lemon. Detecting agreement in multi-party conversational ai. In *Proceedings of the workshop on advancing GROup UNderstanding and robots aDaptive behaviour (GROUND)*, 2023.

[19] Nicolas Spatola, Barbara Kühnlenz, and Gordon Cheng. Perception and Evaluation in Human–Robot Interaction: The Human–Robot Interaction Evaluation Scale (HRIES)—A Multicomponent Approach of Anthropomorphism. *International Journal of Social Robotics*, 13(7):1517–1539, 2021.

[20] SPRING Project. D1.4: User feedback from the preliminary validation (realistic environments). `https://spring-h2020.eu/wp-content/uploads/2022/02/SPRING_D1.4_Preliminary-Experimental-Validation_VFinal_31-01-2022.pdf`.

[21] SPRING Project. D1.5: User feedback from the intermediate validation (realistic/relevant environments). `https://spring-h2020.eu/wp-content/uploads/2023/07/SPRING_D1.5_User-feedback-from-the-intermediate-validation-realistic-relevant-environments_VFinal_31.05.2023.pdf`.

[22] SPRING Project. D5.1: Initial high-level task planner and conversational system prototype for realistic environments. `https://spring-h2020.eu/wp-content/uploads/2021/06/SPRING_D5.1_Initial_High-level_Task_Planner_and_Conversational_System_Prototype_for_Realistic_Environments_vFinal_31.05.2021.pdf`.

[23] SPRING Project. D5.2: Multi-party asr and conversational system in realistic environments. `https://spring-h2020.eu/wp-content/uploads/2023/07/SPRING_D5.2_Multi_party_ASR_and_conversational_system_VFinal_28-06-2022.pdf`.

[24] SPRING Project. D5.3: High-level task planner in relevant environments. `https://spring-h2020.eu/wp-content/uploads/2023/07/SPRING_D5_3_High_Level_task_planner_in_relevant_environments_VFinal_28.03.2023-1.pdf`.

[25] SPRING Project. D5.4: Multi-party conversational system in target environments. `https://spring-h2020.eu/results/`.

[26] SPRING Project. D6.3: Robot non-verbal behaviour system in realistic environments. `https://spring-h2020.eu/wp-content/uploads/2022/02/SPRING_D6.3_Robot_non-verbal_behaviour_system_in_realistic_environments_VFinal_25.01.2022.pdf`.

[27] SPRING Project. D6.3: Robot non-verbal behaviour system in target environments. `https://spring-h2020.eu/results/`.

[28] SPRING Project. WP5: Spoken Conversations. `https://gitlab.inria.fr/spring/wp5_spoken_conversations`.

[29] Stanford Artificial Intelligence Laboratory et al. Robotic operating system. `https://www.ros.org`.

[30] David Traum. Issues in multiparty dialogues. In *Advances in Agent Communication: International Workshop on Agent Communication Languages, ACL 2003, Melbourne, Australia, July 14, 2003. Revised and Invited Papers*, pages 201–211. Springer, 2004.

[31] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

[32] Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. DialogLM: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773, 2022.