THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
l'Institut National des Sciences Appliquées de Toulouse (INSA de Toulouse)

Présentée et soutenue le *Date de défense (29/10/2021)* **par :**

AMANDINE MAYIMA

Endowing the Robot with the Abilities to Control and Evaluate its Contribution to a Human-Robot Joint Action

---

JURY

SILVIA ROSSI	Professeure Associée	Rapporteur
PETER FORD DOMINEY	Directeur de Recherche	Rapporteur
RACHID ALAMI	Directeur de Recherche	Directeur de Thèse
AURÉLIE CLODIC	Ingénieure de Recherche	Directrice de Thèse
SIMON LACROIX	Directeur de Recherche	Membre du Jury
GUY HOFFMAN	Professeur Associé	Membre du Jury
ELISABETH PACHERIE	Directrice de Recherche	Membre du Jury

École doctorale et spécialité :
MITT : Domaine STIC : Intelligence Artificielle
Unité de Recherche :
Laboratoire d'Analyse et d'Architecture des Systèmes (LAAS-CNRS)
Directeur(s) de Thèse :
Rachid ALAMI et *Aurélie CLODIC*
Rapporteurs :
Silvia ROSSI et *Pierre DOMINEY*

Acknowledgments

A faire en dernier :-)

Contents

Acronyms

BDI Belief-Desire-Intention. 3

HATP Hierarchical Agent-based Task Planner. 93

HRI Human Robot Interaction. 79, 83, 86

JAHRVIS Joint-Action based Human-Aware supeRVISor. 3, 66

MuMMER MultiModal Mall Entertainment Robot. 35, 36, 66

REG Referring Expression Generation. 79, 93, 94, 95, 99

SSR Semantic Spatial Representation. 48, 51, 60, 63

Introduction

Human Robot Interaction

Summary of the Thesis

List of Publications

Published

- Sarthou, G., Mayima, A., Buisan, G., Belhassein, K., & Clodic, A. (2021, August). The Director Task: a Psychology-Inspired Task to Assess Cognitive and Interactive Robot Architectures. In *2021 30th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN).*

- Mayima, A., Clodic, A., & Alami, R. (2020, August). Toward a Robot Computing an Online Estimation of the Quality of its Interaction with its Human Partner. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 291-298).

- Singamaneni, P-T., Mayima, A., Sarthou, G., Sallami, Y., Buisan, G., Y., Belhassein, K., Waldhart, J., & Clodic, A. (2020, March). Guiding Task through Route Description in the MuMMER Project. [Video]. In *HRI '20: ACM/IEEE International Conference on Human-Robot Interaction.* (pp.643-643).

- Belhassein, K., Fernández Castro, V., & Mayima, A. (2020, August). A Horizontal Approach to Communication for Human-Robot Joint Action: Towards Situated and Sustainable Robotics. In *International Research Conference Robophilosophy 2020.* (pp.204-214).

- Foster, M.E., Craenen, B., [...], Mayima, A., [...], Lammi, H., & Tammela, A. (2019, November). MuMMER: Socially Intelligent Human-Robot Interaction in Public Spaces. Artificial Intelligence for Human-Robot Interaction. In *Symposium (AI-HRI), AAAI Fall Symposium Series 2019.*

- Mayima, A., Clodic, A., & Alami, R. (2019, November). Evaluation of the Quality of Interaction from the robot point of view in Human-Robot Interactions. In *The 11th International Conference on Social Robotics (ICSR) (1st Edition of Quality of Interaction in Socially Assistive Robots (QISAR) Workshop).*

Accepted

- Mayima, A., Clodic, A., & Alami, R. Towards robots able to measure in real-time the Quality of Interaction. To be published in *International Journal of*

Social Robotics.

Submitted

- Mayima, A., Sarthou, G., Buisan, G., Singamaneni, P-T., Sallami, Y., Belhassein, K., Waldhart, J., Clodic, A., & Alami, R. Direction-giving considered as a Human-Robot Joint Action. Submitted to *User Modeling and User-Adapted Interaction (UMUAI).*

- Belhassein, K., Fernández Castro, V., Mayima, A., Clodic, A., Pacherie, P., Guidetti, M., Alami, R, & Cochet, H. Addressing joint action challenges in HRI: Insights from psychology and philosophy. Submitted to *Acta Psychologica.*

Chapter 1

Human, Robot and Interaction Models: the Funding Principles of a Decision-Making System for Human-Robot Collaboration

Contents

This first chapter aims at setting the context for this thesis. First, we present some related works on human-human and human-robot social interactions. These works nourished the thoughts which led to this manuscript. Then, we develop

key elements for collaboration such as joint action, commitment and shared plans. Finally, we explore Belief-Desire-Intention (BDI) and cognitive robotic architectures which have inspired us to design our own architecture in which, JAHRVIS — the main contribution of this thesis — endows a robot with the abilities not only to control, but also to evaluate its joint action with a human.

1.1 Social interactions

check refs to CA

1.1.1 How to define a social interaction?

First, let's take a look at the dictionary and see how the word "interaction" is defined. According to the Oxford dictionary, an interaction is a "reciprocal action or influence" and more precisely a "communication or direct involvement with someone or something". As for the Cambridge dictionary, it defines it as an occasion when two or more people or things communicate with or react to each other". Those definitions can give an hint about what it an interaction between humans but they are not specific enough. Now, going through social psychology literature, one of the first attempt to define "social interaction" can be found in [Goffman 1967]. Goffman distinguishes three basic interaction units: the social occasion, the gathering and the social situation. The social occasion is an event that is temporally and spatially situated in such a way that it forms a unit that can be looked forward and back upon, by participants that are informed by the event (dinner, meeting, sport game...). The gathering refers to any set of two or more individuals who are at the moment in one another's immediate presence. It can be noted that a social occasion may include several gatherings but that gathering do not need social occasions to occur (they can happen in office spaces, street corners, restaurants. . .). The social situation refers to the full spatial environment that embraces interacting people. It is created as soon as people engage in interaction, when mutual monitoring occurs and ends when the next to the last person leaves. Furthermore, Goffman distinguishes between focused and unfocused interaction (gathering). A focused gathering has its members that can come together to sustain a joint focus of visual and cognitive attention and are open to each other for talk. He calls it encounters or engagements. On the other hand, an unfocused gathering has its members present to one another but not engaged together (e.g. persons waiting for a bus). In this same book, Goffman proposes a definition of social encounter: "an occasion of face-to-face interaction, beginning when individuals recognize that they have moved into one another's immediate presence and ending by an appreciated withdrawal from mutual participation".

A couple of years later, Argyle wrote a book entitled Social Interaction [Argyle 1973], where he lay the foundations the basis to understand social interactions. He came to the view that social interaction could be interpreted as a set of social skills, and that it may therefore be possible to train those skills as

manual skills are trained. For example, during an encounter between two persons, each must be able to perceive the social cues (verbal or non-verbal signals) of the other which are then filtered through the perspective each has acquired through socialization and experience. The interpretation of context and social cues is then applied to arrive at a definition of the situation, which in turn guides both behavior and action.

Then, [Rummel 1976] proposes a definition of a few words: "Social interactions are the acts, actions, or practices of two or more people mutually oriented towards each other's selves. It is behavior that tries to influence or take into account another's subjective experiences or intentions".

Finally, the elements brought here, trying to define what is an interaction and more precisely a social interaction, are chosen among a large amount of work. It is possible to find different definitions, as [Enriquez 2017] precises: "although the term "interaction" is frequently used, it does not have a unique definition but presents semantic divergences".

1.1.2 Structure of a social interaction

Most of the research about interaction and social interaction belongs to the field of social psychology. As for the structure of a social interaction, it is more from the field of Conversation Analysis (CA) which mixes sociology, anthropology, linguistics, speech-communication and psychology. In [Robinson 2012], Robinson makes a review of the work that has been done about "overall structural organization". Most of the time in the literature, overall structural organization is discussed in terms of "the overall structural organization of entire, single occasions of interaction". Then "overall structural organization" term is generally used to talk about one particular (albeit large) unit of interaction. However, many different types of interactional units can have an overall structural organization. For example, in [Schegloff 2011], Schegloff encourages to recognize "'overall structural organization' not as something for the unit 'a single conversation' (or encounter, or session, etc.) alone, but for units like turns, actions and courses of action (like answering or telling), sequences, and who knows what else as well." He also mentioned that every unit of organization should probably have a local organization and a global organization. Here, the term "overall structural organization" is to refer to "the overall structural organization of entire, single occasions of interaction". Robinson tells us that this concept has received relatively little analytic attention and thus is still not well understood [Robinson 2012]. Indeed, research has been more focused on analyzing the organization of individual sequences of action such as turn-takings or conversation openings. Several terms have been used to talk about a "supra-sequential coherence": big package, set of pre-organized sequences, (social) activity, project of activity or plan of action. As the subject has not been investigated so much, it has not evolved a lot in 40 years, since what was proposed by Sacks [Sacks 1995] to define overall structural organization of single occasions of interaction: it "deals, roughly, with beginnings and endings, and how beginnings work to get from begin-

nings to something else, and how, from something else, endings are gotten to. And also the relationship - if there is one - between beginnings and endings". Robinson summarizes research about the subject by saying that single occasions of interaction (in a generic or context-free sense) are normatively organized as: (1) beginning with an opening (2) ending with a closing and (3) having "something" in between opening and closing" which can be referred to as topics.

1.1.2.1 Opening

Openings are used to begin an encounter. One of the main reference on the subject is [Schegloff 1986]. Openings and related issues vary depending on the nature of interactions. For example, opening of a phone call to a family member or a friend will be organized as follow: (1) summons-answer (the one calling talks first) (2) identification/recognition of each other (3) greetings and (4) how-are-you. Whereas, in primary-care medical visits, opening is sequenced as: (1) greeting (2) securing patients' identities (2) retrieving and reviewing patients' records and (4) embodying readiness (sitting down and facing one another). More examples from the literature can be found in (Robinson, 2012).

Another work, [Kendon 1990], focuses on the greeting part, but more precisely the greeting behavior with the associated non-verbal cues. The greeting behavior is divided in three main phases: the distance salutation, the approach and the close salutation. The distance salutation only occurs if the greeters as far enough such as they need to get closer if they wish to continue the interaction. This phase starts after one or both participants sight one another and at least one of them identifies a wish to engage in a greeting. In case one of the participant has not seen the other one, he signals his presence by vocalizing the other one's name or by clearing his throat. Then, they orient their bodies towards each other and exchange glances in a subtle acknowledgement that the greeting is desired by both. During this phase, people can also wave or give a sign with their head (e.g. nod). The approach is divided into two sub-phases: the distant approach (Kendon does not use this term) and the final approach. During the distant approach, people tend to look away whereas when the final approach starts (the greeters are 3 meters or less from one another), they look back at each other and, they smile. Finally, there is the close salutation, the most normalized phase of the greeting. It happens when people are 1,5 meters or less from each others. Then, they can have a non-contact close salutation during which people exchange verbal greetings, or they can hand-shake or embrace (or do something else according to their culture). The greeting is over.

1.1.2.2 Topics

Episodes of interaction vary a lot in their contextualized nature, which leads to a large variety of topics and sequences of topics. Interactions that happen in ordinary or institutional contexts can be pre-organized around one or more topics. Robinson give examples such as an emergency call or an expected call back by a friend to discuss an expected single item of business.

1.1.2.3 Closing

One of the main reference to talk about closings is [Schegloff 1973]. A closing can be divided into two phases: the topic termination and the leave-taking. The topic termination has a pre-closing statement which signals to the partner the wish to close the conversation. Then, the leave-taking follows the pre-closing statement and its response and, includes the goodbye exchange. Finally, the partners break co-presence, *i.e.,* , physically walk apart.[1] In the context of a phone call, Clark and French define this co-presence breaking as the *contact termination* when people hang up.

With regards to non-verbal cues, Knapp *et al.* lists and analyzes them [Knapp 1973]. The more frequent are eye contact breaking, head nodding, leaning toward the partner and positioning in the direction of the way of leaving.

1.2 Human-Robot Social Interactions

Now that we have seen how social interactions look like when happening between humans, we are going to see the different ways the human-robot interaction field divided and categorized interactions.

1.2.1 Short-term Interactions

In [Zheng 2013], they define a "short-term interaction" based on the Unified Theories of Cognition of Newell [Newell 1994]. A short-term interaction corresponds to the "cognitive band" of cognition, during which they focus on individual utterances and speech acts for interactions that last for tens of seconds. They leave aside longer-term interactions that can be in the "rational band" (minutes to hours) or the "social band" (days to months). In [Gaschler 2012], their robot is a bartender then, they define a short-term interaction as being a customer ordering a drink – from the attention request towards the bartender to the closing of interaction by payment and exchange of polite phrases. In [**?**], they use "short-term" to refer to short interactions and that are focused on only one particular communicative objective, avoiding long and complex interactions. In [Sanelli 2017], they give three characteristics to a short-term human-robot interaction: (1) users are not familiar with the robot (2) each interaction happens with a different user (3) interaction is short in time. Then the robot has not memory of past interactions.

1.2.2 Long-term Interactions

A survey [Leite 2013] has been done about long-term human-robot interactions, where long-term means, most of the time, several interactions between the same

[1] It is not explicitly mentioned in [Schegloff 1973] but they precise in a footnote that it would not make sense if the parties remain in co-presence after having being through the closing sequence.

human and robot. They defined four contexts for which social robot[2] for long-term interaction have been designed: health care and therapy, education, work environment and public spaces, and people's homes. For example, Kanda *et al.* performed a field trial at an elementary school in Japan for two months [Kanda 2007]. The children were able to interact with the robot for 32 days in total, during 30 minutes after lunch. The robot could switch between one hundred pre-defined behaviors (*e.g.,* hugging, shaking hand or singing) but not all of them were available during the first interactions with a human. Indeed, they had integrated a pseudo-development mechanism, *i.e.,* the more a child interacts with the robot, the more different behaviours the robot displays to that child. Also, the robot confided personal-themed matters to children who have often interacted with it (*e.g.,* "I don't like the cold"). These abilities allowed the robot to maintain the children's interest even after the first week whereas in a first experiment where the robot's behavior was the same all along the two months, most children stopped to interact with the robot from the second week.

In their discussion part, they raise an interesting question: How Long Should "Long-Term" Be? They found out that some authors consider that two months is a long-term interaction. They also point that some Human-Computer Interaction studies on long-term interaction last five weeks. Finally, the authors of the survey give their point of view, which seems well-thought. They argue that it is more important to look at the number of interaction sessions and the length of these sessions (a five minutes-interaction is different from a one hour-interaction). For them, an interaction can be considered as "long-term" when the user becomes familiarized with the robot to a point that their perception of such robot is not biased by the novelty effect anymore. This definition raises another question: when does user's familiarization with the robot become stable? But it is not discussed here.

1.2.3 Interactions divided in phases

Among works on short-term or long-term interactions, some authors divide interactions in phases which have sometimes similarities with the phases of social interactions described in Section 1.1.

Gockley *et al.* divide an interaction in three phases: greeting, core of the interaction and departure [Gockley 2005]. In the greeting phase, Valerie, the robot receptionist, greets people who might be interested in engaging in conversation. To do so, people are classified into "attentional" states: 1. present (people a bit far and moving): Valerie doesn't pay attention to them 2. attending (people closer): Valerie greets them 3. engages (people next to the desk but on the side): Valerie acknowledges their presence but does not expect input from them 4. interacting (people in front of the keyboard): Valerie prompts them for input if they are not typing. In the core of interaction, either Valerie can tell her (fictive) story or chat. Her story is subjective and evolve over time. It is about her social life, her lounge

[2]actually, some of the robots featured in the survey are not social robots such as a Roomba or the Personal Exploration Rover (PER)

singing career, her therapy business, and her job as a receptionist. Furthermore, Valerie has a chatbot system which is very simple. Finally, inputs from visitors are from a keyboard, for easier control and reliability. Finally, at departure, when a person leaves the "interacting" region, Valerie signals the end of the interaction by saying "goodbye."

In [Kidd 2008], they present a weigh loss coach. It introduces the notion of states of relationship. They are three: initial (for the first few days of interaction), normal, repair. According to the state of relationship, the robot answers/questions/speech will not be the same.

In [Kasap 2012], to each user, corresponds an interaction session. Each session is composed of four dialogue phases: welcome, warm up, teach and farewell. The system has a memory of users and past interactions. In the memory, is recorded the context (initial state and goal), contents (events) and the outcome (goal succeeded or not). A bit similar to the relationship state defined in [Kidd 2008], they define a notion that they call relationship level. It is computed from the emotional interactions from the episodic memory associated to a user. It will influence the mood level and then the facial expression and the speech.

In [Gaschler 2012], they divide the interaction in three phases (or states) but from two different viewpoints, the of the customer and the one of the bartender. From the customer viewpoints, the phases are: (1) attention request towards bartender (2) ordering of one or more beverages, and (3) closing of interaction by payment and exchange of polite phrases. Then, in reaction of each phases, there are the ones from the bartender viewpoint: (1) acknowledging the attention request, (2) serving the ordered drink, and (3) asking for payment. They leave open the possibility to have sub-phases inside phases.

We can also find in [Lee 2012] the notion of structure of interaction: interactions start with the vendor identifying the customer, greeting and engaging in small talk with the customer, engaging in the snack transaction, and then enacting social leave-taking.

1.2.4 Hierarchical interactions

Not only, interactions can be divided in phases but also in levels. In [Dautenhahn 2002] and [Ogden 2001], they define two levels of approach for interactions, a global one and a local one. Both papers are from the same authors but present some small differences in their definitions of the levels. In [Dautenhahn 2002], the global level approach defines a unit of interaction as being relatively large, such as the script for a greeting as described by Kendon in [Kendon 1990]. At this level, an interaction may be seen as a unit similar to a schema or script, in the computer/cognitive science senses of these terms. They name this level of interaction a Global Interactional Unit (GIU). Furthermore, a GIU can be divided in phases, each of which has associated behaviors. Behaviors have meaning and their meaning depends on the phase in which they occur, the context (e.g. a 'wave hello' vs. a 'wave goodbye'). Finally, they discuss the advantages

and drawbacks to describe an interaction at a global level. In [Ogden 2001], they prefer the use of the term "long sequences of interaction" rather than "large units of interaction". In [Dautenhahn 2002], their local level approach is quite similar to the way CA views and analyzes interaction (e.g. adjacency pairs). This interactional structure is a much smaller unit, often as simple as an action and a response to that action. This view of interaction has the advantage of greater flexibility and robustness compared to the globally structured view. Flexibility is a result of the possibility of specifying acts that may occur in many global interactional structures. But, as contextual details are ignored, the ability to assign a specific meaning to an action is lost.

In his thesis, Kuo insists about this flexibility and the re-usability [Kuo 2012]. A lower level of design is more appropriate for reuse. For him, a unit of interaction corresponds to an "interaction cue" (or social cue) that a robot can perceive and act upon or express in an interaction. These cues can be verbal, non-verbal, or a combination of both (multi-modal interaction). A complete episode of interaction should be constructed through composition of interaction cues with some common patterns repeated over the course of the interaction (e.g. awareness of human presence).

1.2.5 Patterns of Interaction

Before talking about design patterns or interaction patterns, Goffman in 1983 [Goffman 1983] argues that human interactions follow a specific "order" and characterized a number of patterns in which people interact, such as how greetings unfold and how people leave an interaction.

In![Kahn 2008], they introduce design patterns, that they will later called interaction patterns in [Kahn 2010], inspired from computer science. They propose rules to follow using them and eight patterns. The two main ideas to retain is that a sequence of patterns has to be well ordered and that patterns can be hierarchical. The 8 patterns: 1. The initial introduction: largely scripted, conventionally-established verbal and behavioral repertoire to recognize the other, inquire politely about the other, engage in some physical acknowledgment (e.g. handshake) 2. Didactic communication: one-way communication of information 3. In motion together: walk together 4. Personal interests and history: sharing of personal interests and history with others 5. Recovering from mistakes: creates the potential for both parties to maintain a social affiliation following the mistake 6. Reciprocal turn-taking in game contextual: taking turns with one another when playing games 7. Physical intimacy: to engage in holding or touching or embracing 8. Claiming unfair treatment or wrongful harms: allows to make claim to its moral standing

Following the same idea and going further, Sauppé and Mutlu [Sauppé 2014] introduces the interaction blocks. Compared to Kahn's work, they offer a pattern language and a tool/environment to design human-robot interaction. To conceive their patterns, they collected and analyzed data from 5 kinds of interaction scenarios: Conversation, Collaboration, Instruction, Interview and Storytelling. Then,

they identified common interaction structures, which served as "design interaction patterns": 1. Introductory monologue: A short introduction can be used to introduce other participants to a scenario by giving an overview of the remainder of the interaction or it can be a greeting for example. 2. Question and Answer: A question is a sentence meant to elicit information from other participants. An answer is the response to a question that aims to satisfy the questioning participant's curiosity. 3. Generic Comment and Personal Comment: A comment is a short statement offering the speaker's opinion. Comments are either generic (e.g., "Wow") or personal (e.g., "I tried that and didn't like it"). 4. Monologue and Generic Comment: A monologue is a longer form of speech during which no response is expected.(e.g. telling of a story). Although monologues expect no response, listeners may occasionally offer unsolicited commentary. 5. Instruction and Action: An instruction is a command offered by one participant to direct the actions of another participant. The proper response to this instruction is often an action, although the action might follow the instruction with a delay depending on whether it is an appropriate time to perform that action 6. Finished Comment: Upon the completion of the goals of the scenario, one or more of the participants will note that the scenario is completed by offering a finished comment. 7. Wait: One pattern implicit in all scenarios involving two or more participants is the wait pattern. Finally, they designed a software to easily implement those patterns in a robot.

In his thesis [Kuo 2012], Kuo criticizes Kahn's work. He says that these patterns involve sequences of interaction cues and should be decomposed to a lower level for detailed design and reuse. He proposes his own patterns: 1. Human presence detection: detect when there is a person who might be interested in 2. Showing interest for interaction: express the robot's awareness of a user's presence around it and its interest and willingness to interact 3. User's attention on the robot: Know when a user is paying attention to the robot in an interaction and its information on its screen 4. User identification by face: Provides the fundamental block for personal service and social interaction by recognising the human counterpart in an interaction He checked the validity of his patterns with the analysis of Problem statement, Context of Use, Interaction Modality, Combination with Other Patterns, Technical Performance and Limitations, User Feedback and User's Perception, Resulting Interactive Behavior.

Finally, Peltason and Wrede also based their work on design patterns from computer science, specifically applied to dialogue here [Peltason 2010]. To name a few of them: Simple action request, Interaction opening, Interaction closing, Clarification. During interaction, the registered patterns are employed in a flexible way by admitting that patterns can be interrupted by other patterns and possibly resumed later which leads to interleaving patterns. By default, simpler patterns are permitted to be nested within temporally extended patterns.

1.3 Around the Joint Action Concept

Often, multiple concepts are addressed when referring to collaborative tasks: collaboration, cooperation, coordination, joint action, joint activity, shared/joint attention, shared/joint intention, shared plan, shared/common/joint goal, (joint) commitment, engagement, mental states, theory of mind, mutual knowledge... However, many terms and definitions, whether inside a field[3] or between fields do not reach a consensus. This can be quite confusing, especially for roboticists for which it is initially not the range of expertise. Thus, we will first give an overview of the more characteristic definitions of what is Joint Action. Then, we will present a non-exhaustive set of notions related to Joint Action.

1.3.1 How to define Joint Action ?

An important number of social interactions and encounters are encompassed by the notion of joint action. Broadly considered, joint action is any form of social interaction whereby two agents or more coordinate their actions in order to pursue a joint goal. However, the notion of joint action has particularly been subject to debate in philosophy and psychology. For instance, according to Sebanz *et al.* [Sebanz 2006, p. 70], "joint action can be regarded as any form of social interaction whereby two or more individuals coordinate their actions in space and time to bring about a change in the environment."; while other authors [Carpenter 2009, Cohen 1991, Fiebich 2013, Tomasello 2005, Pacherie 2012] resist the idea that instances of mere coordination – *e.g.,* two partners walking side by side – constitute a joint action, considering that it requires some necessary conditions like sharing goals and intentions.

Moreover, while the notion of joint action is used interchangeably with the notion of collaboration or cooperation for some authors such as Becchio *et al.* [Becchio 2010] and Kobayashi *et al.* [Kobayashi 2018], other authors establish a hierarchy of interactions depending on the processes involved [Amici 2015, Chalmeau 1995]. According to Amici and Bietti [Amici 2015], for example, coordination is a fast low-level process of behavioral matching and interactional synchrony which could, but not necessarily, facilitate middle-level processes like cooperation (where some individuals bear certain costs to provide benefits to others) or high-level processes like joint action, which requires other resources like turn-taking and alignment of linguistic resources during dialogue.

If we look at Sebanz and colleagues definition of joint action, it could be considered as a kind of activity based on the usual sense of this term. Thus, some authors use the concept "joint activity" interchangeably with "joint action" [Tollefsen 2005, Gräfenhain 2013] while others see the joint activity composed of joint actions [Clark 1996, Klein 2005]. Clark says that "Joint activities advance mostly through joint actions". He defines the properties of a joint activity among which there are: it is carried out by 2 or more participants, each participant has

[3] Here, philosophy, psychology or robotics

a public role or they try to establish and achieve joint goals, and they may have private goals. He also highlights the need for coordination: "What makes an action a joint one, ultimately, is the coordination of individual actions by two or more people. There is coordination of both *content*, what the participants intend to do, and *processes*, the physical and mental systems they recruit in carrying out those intentions" [Clark 1996, p .59].

In this manuscript, the word joint action will be used to indifferently refer to an activity/task composed of several (joint) actions, *i.e.,* a high level joint action or as a single action, but in both cases it will imply that it is a "social interaction social interaction where two or more individuals coordinate their actions in pursuit of a common goal" [Castro 2020].

1.3.2 Two possible segmentations around Joint Action

Before going through the mechanisms involved in joint action, we will briefly present two segmentations of joint action: a temporal segmentations, *i.e.,* the different phases a joint action goes through, and a cognitive model for human agency *i.e.,* the different neurocognitive levels that are involved in joint action. It seemed necessary to present these two segmentations as the processes related to joint action described in Section 1.3.3 are sometimes involved in one phase/level but not in another. However, we will not go through these details as it is not necessarily relevant to the rest of the manuscript.

1.3.2.1 Temporal segmentation of Joint Action

As a joint action is a form of social interaction, it can also be divided in three phases as presented in Section 1.1.2: an initiation, a body and a closing [Heesen 2017]. Each phase a role; the initial phase establishes among other things the joint commitment, *i.e.,* who is to participate, in what roles, what actions will be performed, and when and where they will be performed different roles for the participants [Clark 2006]. Then, in the body they coordinate to perform their goal. Finally, "to complete a joint action, participants first need to arrive at the mutual conviction that they are both indeed ready to terminate it" [Heesen 2017], if they achieved their goal for example.

1.3.2.2 Neurocognitive segmentation of Joint Action

To describe the neurocognitive mechanisms involved in joint action, we will base ourselves on the conceptual framework established by Pacherie [Pacherie 2008]. This framework is particularly relevant for the rest of manuscript as it has a lot similarities with the three-layered robotic architecture that will be described in Section 2.3, as noticed in [Clodic 2017].

1.3.3 What is necessary for Joint Action ?

Leaving aside the debate on the concept of joint action, we aim to focus on the mechanisms that enable the consecution of joint actions. What we found to be the mechanism on which every author (or almost) agrees on to say that it is required for a joint action is the *coordination*. This mechanism itself is supported by other cognitive and sensorimotor processes. Also, philosophers introduced another concept involved in joint action which is the *shared intentions*. In Figure 1.1

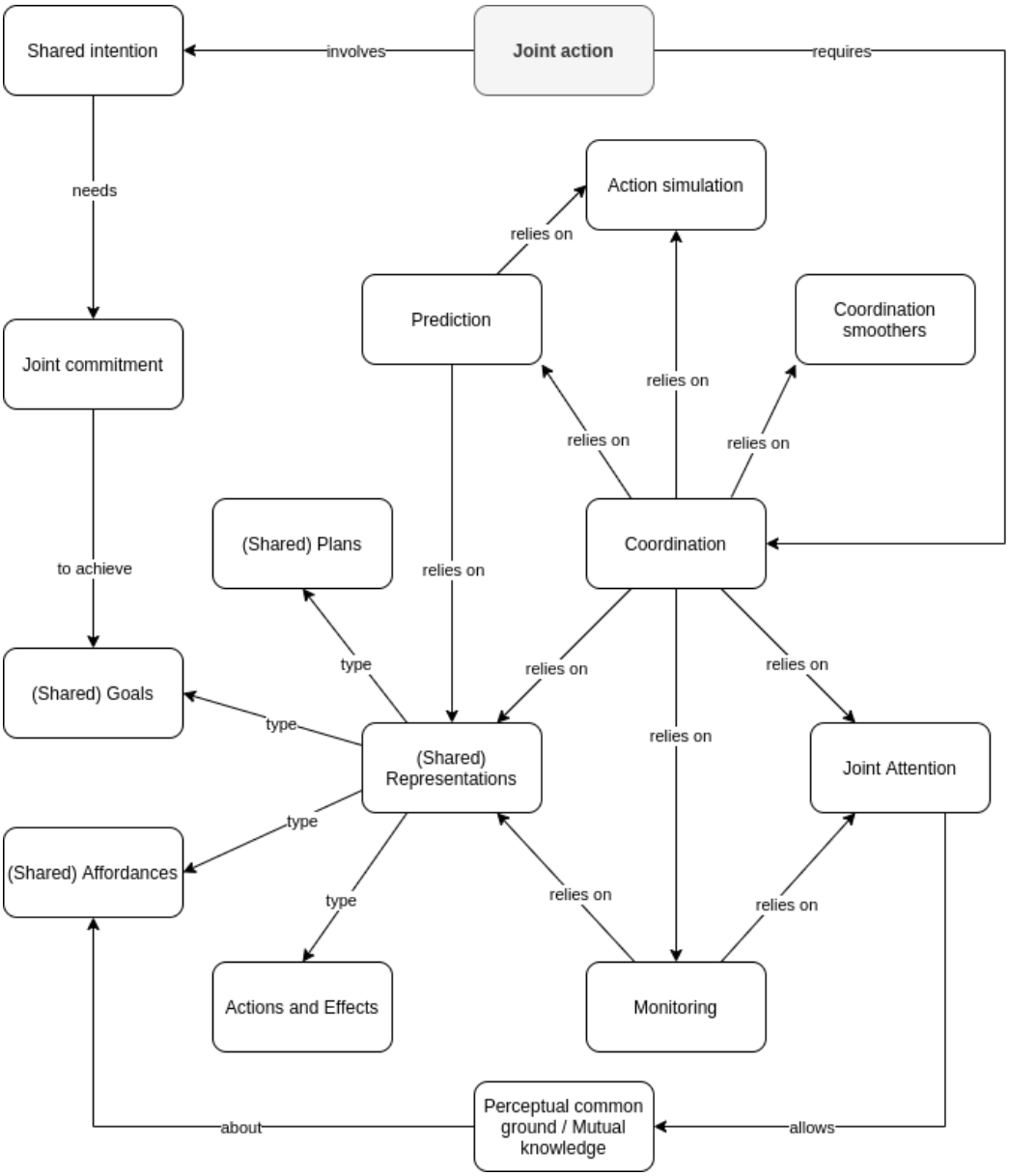


Figure 1.1: Overview of a non-exhaustive set of processes related to joint action

Three interrelated mechanisms appear to be key conditions for joint action:

coordination, planning, and motivational alignment, each of them being supported by other processes. There has been an important deal of conceptual and empirical work investigating these processes [Knoblich 2011, Pacherie 2012], from the sharing of a common ground to the anticipation of a partner's actions by way of emergent coordination [Curioni 2019].

Most of all, joint actions require individuals to anchor their plans in the actual situation and generate particular coordinated actions [Knoblich 2011, Vesper 2011]. This coordination can rely on different mechanisms, which are not necessarily intentional, including for example perception-action matching [Brass 2001], perception of joint affordances [Ramenzoni 2008], joint attention [Sebanz 2006]. As for intentional coordination – sometimes referred to as planned coordination [Curioni 2019] –, it requires the partners: (i) to represent their own and others' actions, as well as the consequences of these actions, (ii) to represent the hierarchy of sub-goals and sub-tasks of the plan, (iii) to generate predictions of their joint actions, and (iv) to monitor the progress toward the joint goal in order to possibly compensate or help others to achieve their contributions [Pacherie 2012].

Indeed, joint action often involves planning several aspects, which involves the representation of the goal and the whole plan, and/or even the sequence of actions to be performed. The formation of these types of representations could rely on different mechanisms, which include high-level processes such as theory of mind, team reasoning, or verbal negotiation [Bratman 2013] or more low-level processes as minimally representing the joint action goal and knowing that it will be achieved with others [Vesper 2010]. An example of the mechanisms involved in planning a joint action is task co-representations, which allows individuals to represent the details of each other's task.

These representations allow individuals to generate predictions about the other's actions, which in turn facilitate the adjustment and coordination between the partners. Interestingly, individuals can also facilitate the others' and their own predictions by communicating relevant and reliable information for joint action. The objective is to make actions more transparent and predictable so that the decision-making on the interaction can be fluent and successful. Thus, participants in joint action often negotiate on the fly the sub-tasks, sub-goals, or ways to proceed regarding the collective task through different explicit exchanges [Clark 1996].

1.4 Human-Robot Joint Action

part on emotions

1.4.1 Focus on Communication in Human-Robot in Joint Action

Cohen and Levesque, Teamwork, p 490

1.4.2 Focus on Repairs in Human-Robot Joint Action

One last subject of CA could be interesting for our problematic in human-robot interaction, it is the repair. It is a mean of correcting a misunderstanding or a mistake in an interaction, or of correcting a deviation from the normal rules of interaction. The ability to engage in repair is essential in interaction: errors and misunderstandings are likely to arise and must be corrected if the goal of the interaction is to be successful. Generally, they are classified in four categories [Schegloff 1977, Wooffitt 2008]:

- Self-initiated self-repair: Repair is both initiated and carried out by the responsible of the trouble
- Other-initiated self-repair: The responsible of the trouble takes care of the repair himself but the trouble have been pointed out by the other
- Self-initiated other-repair: The responsible of the trouble signals that a repair is needed and get the other one to repair (e.g. he forgot a name and asks for help to remember)
- Other-initiated other-repair: The one not responsible of the trouble initiates and carries out the repair. This is closest to what is conventionally understood as "correction".

1.4.3 Focus on Commitments in Human-Robot in Joint Action

1.5 Supervision systems for Decision and execution

1.5.1 BDI architectures

There are several kind of integrated cognitive architectures, a survey describes the most known and representative of them [Chong 2007]. Some of them have their roots in classical artificial intelligence as the Soar architecture. Others take their inspiration in cognitive psychology as ICARUS which tries to produce artificial intelligence mimicking human cognition or the BDI architecture which is based on the studies of folk psychology and intentional systems. Furthermore, inspired by neurology, there is the subsumption architecture. It is a reactive architecture therefore, it is not based on symbolic mental representations but on sensory information fusion to select actions to execute. Finally, at the crossroads between AI, cognitive psychology and neurology, we can find the ACT-Rand CLARIONmodels. The BDI model is based on the philosophical model of human practical reasoning developed and designed by Michael Bratman [Bratman 1987, Bratman 1988]. It has 3 main concepts:

ref

ref

ref

- Beliefs: They are a representation of the agent's knowledge about the world. "[They] can be viewed as the informative component of system state" [Rao 1995]. It is not the word "knowledge" that has been chosen to

define this concept because what the agent perceives of the environment is in fact the likely state of the environment. There is no certainty, its sensors are not accurate or could malfunction. This way of distinguish knowledge and beliefs is one that can be found in the literature of distributed computing [Lamarre 1994].

- Desires[4]: They are a representation of the motivational state of the system. They provide "information about the objectives to be accomplished or, more generally, what priorities or payoffs are associated with the various current objectives" [Rao 1995].
- Intentions: They are a representation of the currently chosen course of action (plan). It is the deliberative component of the system. The selected course(s) of action are determined with a deliberative function, according to the beliefs and desires [Rao 1995].

1.5.2 Supervision for Human-Robot Interactions

1.6 Limitations and Challenges

[4]In one of the first implementation, PRS, "Goals" notion was used instead of "Desires" [Georgeff 1989], then they use it in a interchangeable way in [Georgeff 1991] and finally choose "Desires" [Rao 1995] with the definition given in the AI literature, *e.g.,* desires can be many at any instant and may be mutually incompatible. Therefore, a goal will be a chosen desire [Cohen 1990] and concurrent goals are consistent.

Chapter 2

Joint-Action based Human-Aware supeRVISor: JAHRVIS

Contents

2.1 Introduction

2.2 Representation of a H-R collaborative activity

It is possible to describe and decompose a Human-Robot collaborative activity in various ways. For all the following definitions, we place ourselves in the context of one-to-one human-robot interactions, however we believe that the scheme can be extended to multi-human multi-robot contexts. We draw our inspiration from the literature of sociology and robotics to define a model of interaction with three layered levels: interaction session, tasks and actions; as illustrated in Fig. 2.1. We chose to represent collaborative tasks and their decomposition using the Hierarchical Task Network (HTN) [Ghallab 2016] representation which is often used in cognitive robotics [Ingrand 2017, Lallement 2014] and because it allows to deal with goal-based and situation-based activities at different levels of hierarchy such as task, subtasks and actions and consequently to consider different level of granularity. In the example of a task with an overall bad QoI, it would be interesting to know that in fact it is only a particular action or subtask ruining it. Indeed, the other parts of the task can be ok, or on the opposite, a particular subtask or action can have performed very well among the others. We need and use this granularity also on three levels defined (interaction session, tasks and actions) to finely evaluate the Quality of Interaction, as a task can be of poor quality but the session is globally going well.

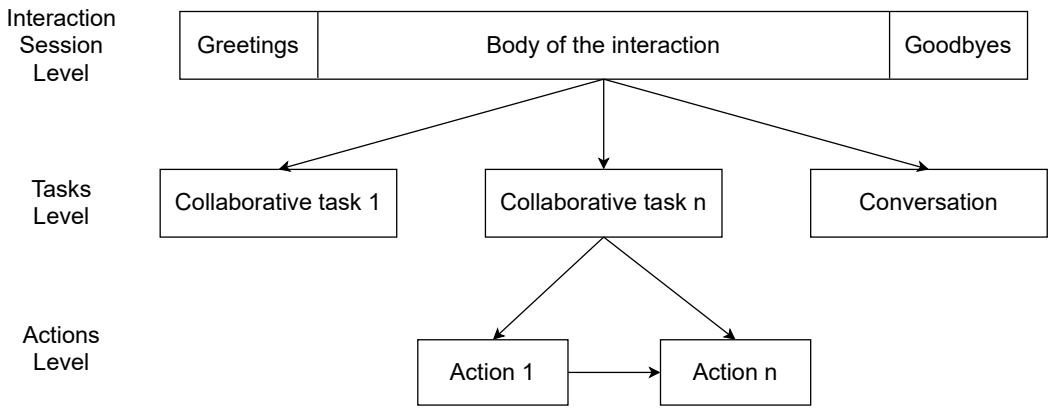


Figure 2.1: The hierarchical structure of an interaction session. The highest level is the interaction session. The second level is composed of the tasks. They are included in the body of interaction of the interaction session and, two types of tasks are considered and may overlap, collaborative and conversational tasks. With this representation, a task can be recursively refined as subtasks until reaching the last level, the actions level, which is considered as atomic. Subtasks are not considered as a "real" level of the interaction session, specially to evaluate the QoI, as it may exist or not according to the task.

2.2.1 Representation of a H-R Interaction Session

We define an **interaction session** as the period during which the robot and a human interact together and are engaged. It is divided in three parts, following the structure proposed by Robinson [Robinson 2012] and the engagement model of Sidner and Lee [Sidner 2003]: the greetings, the body of the interaction and the goodbyes. First, *the greetings* corresponds to the period where an agent starts an interaction by initiating it with another agent. The interaction session lasts as long as the interactants are maintaining the interaction through conversation and collaborative tasks performance which corresponds to the *body of interaction*. Finally it ends when at least one of the interactants is disengaged, either by abruptly ending the interaction or by closing the interaction as described by Schegloff and Sacks [Schegloff 1973], it corresponds to "the goodbyes". For example, for an entertainment robot in a mall, an *interaction session* starts when a person signals to the robot that they want to engaged, by greeting it or by approaching it and looking at it. The body of interaction is composed of conversation and eventually direction-giving tasks and, the session lasts until the person says goodbye or leaves. This is the nominal case and, the duty of the robot is to contribute to maintain the session alive until the human decide to close it. However, in some (extreme) cases, the robot might decide to close the interaction by itself.

Social interactions and collaborative tasks involve engagement. There is no unique definition of what it means to be engaged. We chose one that is frequently used and has been proposed by Sidner and Lee [Sidner 2003]: "Engagement is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake". The robot must be able to exhibit its engagement and disengagement and also to assess them with respect to its human partner. We defined three states for the body of interaction, corresponding to what is happening during the latter: conversation (i.e. a social chit-chat or a goal negotiation, without any physical action performed except communicative gestures), collaborative task (i.e. both agents executing actions in order to achieve a shared goal) or idle phases (i.e. the agents are not chatting or performing a collaborative task together but remain engaged in the interaction session, it happens in-between active interaction phases). For each of these three states, the way to exhibit the engagement varies (e.g. in a conversation, an agent looking at their partner displays their engagement; during a task, an agent correctly performing their action is a way to demonstrate their engagement). That is why there is a need to define what behavior the robot has to exhibit in each state and what behavior it should expect from the human in each state, as these behaviors are usually very specific (e.g. in a direction-giving task, the robot keeps its head oriented toward its partner's face to demonstrate its engagement in conversation and idle contexts and when it gives a direction it expects the human to look at the direction it is showing; in a stack task, when the robot gives an instruction it expects the human to take a given cube).

2.2.2 Collaborative Tasks, Substasks and Actions

Tasks compose the body of the interaction of an interaction session as shown in Fig. 2.1. We distinguish conversation (i.e. agents engage in dialogue to exchange ideas, to ask questions, and to resolve differences) from collaborative tasks (i.e. agents work as partners, collaborating to perform tasks and to achieve common goals). We will not develop more on conversation since it is not the main focus of this paper, assessing the QoI of social dialog being another work.

In collaborative tasks, the robot and the human are committed to achieve a goal together, involving joint actions and shared plans [Grosz 1996]. When a human and a robot perform a task together, as described by Bauer *et al.* [Bauer 2008], we could say that the robot has the intent to help the human, so the human's intention becomes its own intention. Then, they have the joint intention to reach a common goal and, as shown by Michael and Salice [Michael 2017], they have a commitment to the joint activity, leading to perform joint actions. Therefore, during its evaluation and decision-making processes, the robot has to take into account that the human and itself should remain engaged all along an interaction session for the tasks to be successful and both have to manage and contribute to maintain expectations about what the other is doing.

The elements composing a *task* are: a goal, a plan and involved agents. A plan is needed to realize a goal. There are many ways to generate a plan. But no matter the way (using a planner to anticipate execution or relying on a reactive planning scheme), a plan is a sequence of ***subtasks*** which are sequences of actions – *subtasks* are not considered as a "real" level of the interaction session, specially to evaluate the QoI, as it may exist or not according to the task.

Actions are the elementary items of tasks manipulated by the high-level robot supervision controller. They cannot be decomposed further by it (e.g. placement and motion planning are achieved by a lower control system not described here). It is usual to describe an action with its preconditions, its effects and, the agents and entities implied in its execution (e.g. in plans written in PDDL (Planning Domain Definition Language) [Ghallab 1998]). We add to this description the notion of expected reactions (which can themselves be actions) from the other agents once the action is executed.

In our model, an agent (human or robot) is a contributor to the task and has a mental state as described by Devin *et al.* [Devin 2016]. The mental state is a set of facts representing, from the agent point of view, the current world state, the state of the goal and the current task state. Since we are interested here by the robot situation assessment and decisional processes, the mental state of the human is built and managed by the robot as an estimation of the beliefs of the human [Milliez 2014a, Hiatt 2017, Tabrez 2020].

2.5.1 Introduction

Robots dedicated to Human-Robot interactions are not just machines receiving commands and executing them. They should be decisional agents with high-level goals, taking decisions (potentially taking into account social norms), and acting and reacting to not only their actions but those of other agents. Cognitive and interactive robots are becoming more and more capable thanks to the use of human-aware models and algorithms [Kruse 2013, Thomaz 2016], with roboticists endowing them with the ability to execute their share of the work while adapting to contingencies, particularly those caused by human's behaviours and decisions [Hoffman 2007, Baraglia 2017, Lemaignan 2017b]. The decision-making process is based on a range of knowledge about the environment, the interaction,

the context... Nevertheless, curiously and interestingly, very little has been done to allow the robot, while performing its collaborative or assistive activity, to permanently evaluate if things are going well or not, as humans do. We name this ability "the measure of the Quality of Interaction from the robot point of view". We believe that enriching the robot knowledge with a good estimation about how the interaction is going, could enhance its decision-making process and thus, its social behaviour.

For example, if the robot detects that the QoI starts to drop, it can take a decision based on this information and act to try to improve the interaction quality (e.g. it can choose to change some modalities such as the language in which it communicates with the human, the volume of its speakers, or the parameters of its planners). On the contrary, when the QoI is high, the robot can decide to just continue the interaction as planned. Then, endowed with a QoI Evaluator, a robot becomes more adaptive and performs better. Also, a very poor performance all along a task could allow the robot to assess that the human is not really engaged in the interaction, or even is trying to play the robot. In such situation, the robot might perhaps better disengage. Finally, from a methodological point of view, a robot deployed in the wild able to assess interactions, has an asset compared to others as it could reduce the investment in material and human resources to perform user studies. And, a developer might use the logs to improve their design.

In this paper, we only focus on the Quality of Interaction evaluation process and not on how to use its result for decision making. Therefore, we present in the sequel the methods and tools we developed, allowing the robot to evaluate in real-time the quality of the human-robot collaborative activity it is involved in. It is based on a set of metrics we have defined, focused on two concepts: the measure of human engagement and the measure of the effectiveness of collaborative tasks performance. However, this is by no means exhaustive, and other metrics and parameters could (and should) be added later. Our work can be seen as a toolbox among which it is possible to pick the desired metrics according to tasks or contexts. We propose a way to aggregate these metrics, producing the QoI. The evaluation of the QoI is performed at three different levels of abstraction: the interaction session level, the task level and the action level. In further work, this ability could provide additional information to the robot and open the possibility for reconsidering its behaviour in case it estimates that the quality of the interaction is degrading (e.g. changing its plan or the way it is achieving it, informing the human or requesting a change in their behaviour, or even deciding to disengage).

The chapter is organized as follows. In the next section, we briefly discuss related work and the main challenges. In section 2.2 we present the representation of human-robot collaborative activity which we use and its hierarchical decomposition. Finally, in sections 2.5.3 and 2.5.4, we introduce our concept and proposed set of metrics to evaluate the Quality of Interaction.

2.5.2 Related work

Inspired from the evaluation methods used in Human-Computer Interactions and User Experience fields, the field of Human-Robot Interaction (HRI) has elaborated its own methods to evaluate robotic systems when they interact with humans. There are various ways to evaluate a human-robot interaction from the human perspective. Bethel *et al.* [Bethel 2010] divided them into five categories: (1) self-assessments, (2) interviews, (3) behavioral measures, (4) psychophysiology measures, and (5) task performance metrics. They reviewed metrics used for each of the categories. They can be grouped into two types: (1) and (2) are subjective metrics and, (3), (4) and (5) are objective ones. Since our aim is to have a robot able to evaluate interactions by itself, human subjective metrics are not usable. Then we focused on the study of existing objective metrics meant to measure how the interaction goes. Steinfeld *et al.* [Steinfeld 2006] proposed a set of metrics to be used in a wide range of tasks whose goal is to assess the system performance by measuring the task effectiveness (i.e., how well the task is completed) and the task efficiency (i.e., the time required to complete a task). Their work is very thorough and inspiring but does not target the evaluation of the quality of an on-going interaction. Hoffman [Hoffman 2019] defined a type of quality of interaction, the *fluency*, pointing out that the notion is not well defined and somewhat vague but can still be assessed and recognized when compared to non-fluent scenario. To measure it, they propose a list of objective metrics, only based on duration measures, designed to be quite general: robot idle time, human idle time, concurrent activity (i.e., active time of both the robot and the human), functional delay (i.e., time difference between the end of one agent's task and the beginning of the other agent's task). It is an interesting way to measure the fluency and thus the quality of the human-robot interaction but it only applies to shared workspace tasks and is dedicated to an offline evaluation.

Systems targeting real-time measurements during human-robot interactions, with the purpose to "close the loop" and use the information for decision-making, have been developed. Tanevska *et al.* [Tanevska 2017] proposed a framework allowing the robot to perceive with face detection and evaluate in real-time the affective state (i.e. anger, happiness, sadness, surprise, etc) and the engagement state (i.e. whether the person is interested or bored in the interaction) of the people it is interacting with. However, the human affective state measure might not be enough to assess an interaction or a task as an affective state is actually a facial expression which can be misinterpreted (e.g. a smile can be a sign of happiness or embarrassment) and which might be not visible when one of the agent perform an action and looks somewhere else. Moreover, as the notion of engagement is very task specific, it needs further exploration. Real-time engagement measurement has also been investigated by Anzalone *et al.* [Anzalone 2015] using metrics such as gaze, head pose, body pose and response times. Their work is interesting and could be an element among others to assess the interaction quality but, it is dedicated to face-to-face interactions.

Cameras are not the only sensor used to assess interactions on-the-fly, some use

human physiological responses such as skin conductance and temperature, heart or brain signals. Itoh *et al.* [Itoh 2006], Bekele *et al.* [Bekele 2014] or Kulic *et al.* [Kulic 2007] use them to detect human affective states such as anxiety or liking in real-time. However, physiological measures often imply a lot of sensors which can be invasive for the human. And, as explained by Kulic and Croft [Kulić 2003], physiological signals may be difficult to interpret and there is a large variability in physiological response from person to person. Thus, it can be difficult for a controller to determine which emotional state the subject is in, or whether the response was caused by an action of the system, or by an external stimulus. Moreover, we claim that the human affective state only is not enough to assess the quality of an interaction, a human could be satisfied with an interaction or a task result even though they were stressed during it.

Finally, Bensch *et al.* [Bensch. 2017] proposed a formal approach to compute interaction quality in real-time. Their work focused on how to combine metrics together which is in the same line as ours. However, they do not provide implementation examples, remaining at an abstract level.

In summary, while a substantial number of studies have been devoted to the evaluation of collaborative interactions for analysis purposes once the interaction is over, there is a lack of methods allowing the robot to evaluate in real-time the quality of the interaction based on multiple metrics and not only anxiety or engagement. We claim that such an ability is very important and should strongly influence the situation assessment as well as the decisional abilities of interactive and collaborative robots.

2.5.3 The Quality of Interaction (QoI)

We believe the real-time assessment of the Quality of Interaction (QoI) with a human partner (i.e. what the robot "thinks" about how the interaction is going) is a new knowledge that could enhance the robot decision-making process. We define the Quality of Interaction as a measure that indicates how good is the interaction during human-robot collaborative activities. It is computed in real-time based on a set of metrics, at three different levels: the interaction session level, the tasks level and the actions level. The QoI of a given level is computed from selected metrics but also from the QoIs of the level below as shown in Fig. 2.2.

The QoI of each level is computed as a score between [(1) for a good quality] and [(−1) for a poor one]. Metrics used to compute the QoI are divided in three categories:

- $Mp \in [0, 1]$ if it can only have a positive effect on the evaluation;
- $Mn \in [-1, 0]$ if a metric can only have a negative effect on the evaluation;
- $M \in [-1, 1]$ if a metric can have a positive or a negative effect.

Defined by the designer according to the needs and context, a metric can belong to one category or another depending on the target application. When needed,

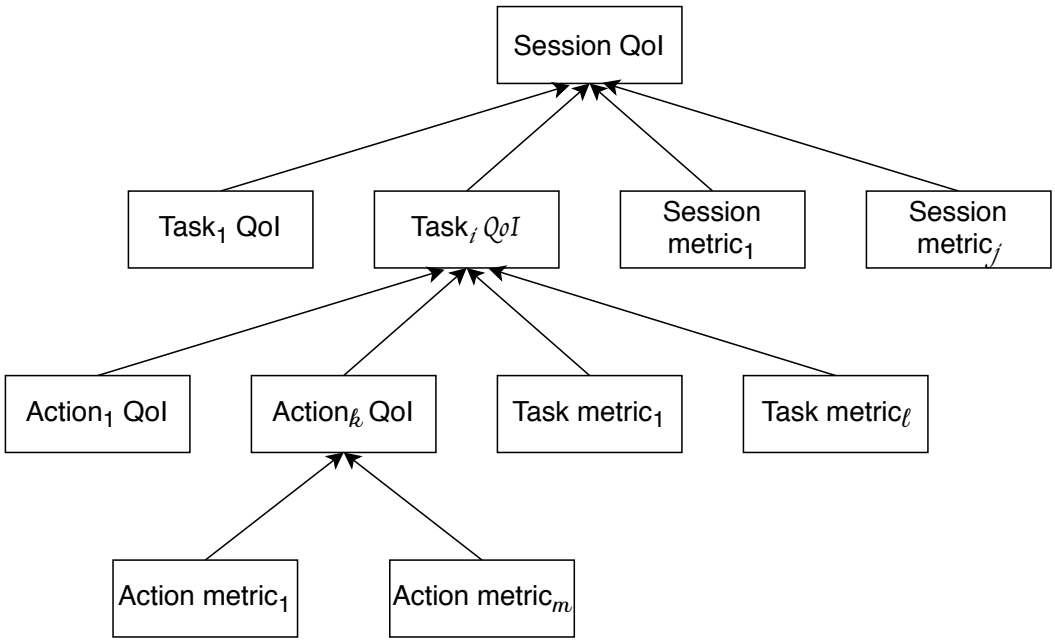


Figure 2.2: Representation of the QoI dependencies, with i the number of performed tasks during the interaction session, k the number of performed actions during the task i, j the number of metrics to measure the interaction session QoI, l the number of metrics to measure the task i QoI and m the number of metrics to measure the action k QoI.

metrics values are scaled with the equations presented in Appendix A.

The evaluation of the Quality of Interaction at the level $l \in \{session_f, task_j, action_k\}$ (with f, j and k respectively the identifiers of a given interaction session, task and action), QoI_l, is computed with:

$$QoI_l = \frac{\sum_{i=0}^{x} W_i * M_i}{\sum_{i=0}^{x} W_i} + A * \frac{\sum_{i=0}^{y} Wn_i * Mn_i + \sum_{i=0}^{z} Wp_i * Mp_i}{\sum_{i=0}^{y} Wn_i + \sum_{i=0}^{z} Wp_i} \tag{2.1}$$

with W_i, Wp_i, Wn_i respectively the corresponding designer-set weights of M_i, Mp_i, Mn_i, A the designer-set weight of the right part of the + sign and x, y, z respectively the number of the metrics M_i, Mp_i, Mn_i.

Equation 2.1 aggregates the values of the metrics chosen to be indicators of the interaction level quality. As all metrics do not have the same importance in the measure of the QoI, each of them is weighted. Values of these weights are empirically defined. There are two parts in the equation, the left part of the + sign and the right part. The left part of the + sign is a weighted mean of the third category of metrics, the M metrics. The right part is a weighted mean of the metrics seen as bonus (i.e. Mp metrics) or penalty (i.e. Mn metrics). This latter part is weighted

with A – whose value is also empirically[1] defined – to be able to adjust its influence on the left part. In such a way, if there are no Mn metrics to compensate for the Mp metrics, it is possible to limit the positive influence of the Mp metrics on the M metrics with A. It is the same if there are no Mp metrics, A can compensate the impact of the Mn metrics on the M metrics. Even though $M, Mp, Mn \in [-1, 1]$, the final result of QoI_l might be less than -1 or greater than 1 because of the addition of the M with the Mn and Mp. If it happens, QoI_l minimal value is set to -1 and its maximal value is set to 1.

2.5.4 A set of metrics

In this section, we present a few measures to assess the QoI of an interaction session in Sect. 2.5.4.1. Then, we present metrics for the different levels based on engagement in Sect. 2.5.4.2 and effectiveness estimations during human-robot joint activities in Sect. 2.5.4.2. For example, if the human is engaged and if tasks are performed effectively, the QoI will tend to be high and *vice versa*. Both concepts are difficult to measure, so we do not exactly measure them but we compute their trends from the set of metrics presented in this section. This set is not exhaustive and will be extended in future work but it gave promising results as we show with our implementation in Chapter 3. All metrics are meant to be used for online evaluations of interactions. They are summarized in Table 2.1.

2.5.4.1 Measures to assess the QoI at the interaction session level

According to the context, the duration of an interaction session can be an indicator of the human engagement. Indeed, a human leaving only a few seconds after the beginning of the interaction is probably less engaged than a human staying with the robot several minutes. Also depending on the context, the number of executed tasks is a measure which can be considered as interesting information with respect to the engagement of the human, as well as the ratio of successful tasks. The more the human executes successful tasks with the robot, the higher the session QoI might be. Finally, it can be valuable to take into account how the session has been terminated in the evaluation of the quality of an interaction session. For instance, the fact that the human leaves abruptly in the middle of a task, during an idle time or a conversation without saying goodbye, or only at an appropriate time saying farewell to the robot is significant in terms of social interaction quality.

2.5.4.2 Metrics related to human engagement

Michael *et al.* [Michael 2016] stated that commitments[2] facilitates "the planning and coordination of joint actions involving multiple agents. Moreover, commitment

[1] Values are empirically defined given intuition regarding the importance of a given metrics for a given task and a set of testing experiments

[2] In the robotic domain, it is the word "engagement" and not "commitment" which is often used, unlike in the psychological and philosophical fields.

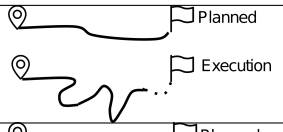
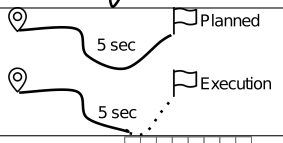
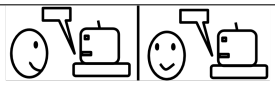
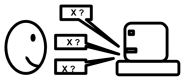
	Metric names		Measures	Illustration	Session	Task	Action
Effectiveness	Progress towards goal	Distance-to-Goal	Geometric distance			x	x
		Time-to-Goal	Time			x	x
		Steps-to-Goal	Number of executed actions/substasks			x	
	Deviation from standard duration		Time			x	x
Engagement	Fulfilling robot expectations about social interaction		e.g. attention ratio, with-me-ness,...		x	x	x
	Human contribution to the goal		e.g. number of repeated instructions, number of successful human actions,...			x	x

Table 2.1: The set of metrics presented in Section 2.5.4.

also facilitates cooperation by making individuals willing to contribute to joint actions to which they would not be willing to contribute if they, and others, were not committed to doing so". As it is an important element of the joint action, we want to provide the robot with a way to estimate the engagement of its partner during an interaction.

Metrics allowing to state if an agent is engaged or not in an interaction are often specific to the type of interaction. For example, Fan *et al.* [Fan 2017] implemented their measure of the human engagement as a kind of hysteresis: when the human gaze is on the robot, they are considered as engaged and when the human gaze is somewhere else during more than 3 consecutive seconds, they are considered as not-engaged.

In the same vein, we think that the measure of the engagement for a collaborative activity can be divided in 2 types of metrics, summed up in Table 2.1: the Human contribution to the goal and the Fulfilling robot expectations about social interaction.

We define in this section examples of metrics of each types which can be used to estimate the level of engagement of the human partner.

Human contribution to the goal A good and very promising indicator could be the ability from the robot to evaluate how well the human actions help to the

goal progression. We call this indicator *Human contribution to the goal*. To the best of our knowledge, there is no general method to estimate it.

As a first version of the *Human contribution to the goal*, we chose to measure it through the number of times the robot has to repeat an instruction or a question before the human performs correctly, when it expects the human to answer or to perform the action. As, if it needs to repeat, it means that the human is not correctly contributing to the goal, intentionally or not, as they are not performing their part of the HR action as they should. The more the robot needs to repeat because of the human's bad performance, the less they are contributing to the goal, the more the action QoI should decrease.

Fulfilling robot expectations about social interaction During a social interaction, agents are expected to behave in a certain way and so the robot has expectations about the human. Then, the robot can monitor the human behavior to check if they are acting as they are expected to. For example, most of the time, when the robot speaks to the human, it will expect them to look at it and so it can monitor if it is the case or not as implemented by Fan *et al.* [Fan 2017]. Quite similarly, Lemaignan *et al.* [Lemaignan 2016] developed a way to measure if the human is *with* the robot during their interaction, based on attention assessment, by computing if the human is looking at the desired attentional target or not. This latter metric will be integrated to our framework in future work.

As the works of Lemaignan *et al.* and Fan *et al.*, we estimate the *Fulfilling robot expectations about social interaction* with the human head orientation, in the context of our implementation described in Chapter 3. We compute an attention ratio i.e., the time during which the human is attentive to the robot (i.e. staying close enough and looking at it) when it speaks compared to the total time of the speech:

$$Ar = \frac{duration_{isAttentiveTo(robot)=true}}{duration_{robot\_speaks}} \tag{2.2}$$

Metrics related to effectiveness One can elaborate metrics to measure how well a task or an action is achieved. As discussed by Olsen and Goodrich [Olsen 2003], there are a variety of metrics such as time-based metrics which reward the speed of performance or the response times; error metrics which are based on counting retrials, failures, or mistakes; coverage metrics which measure to what extent a goal is achieved, as well as other possible metrics. We use some of them such as counting retrials, however these metrics alone were not enough for our example task as we are in a HRI context.

One can measure for different kinds of tasks, the ratio of successful[3] executions to the total number of executions (e.g. $R = \frac{Succ}{Exec}$) or the deviation from the initial

[3]Obviously, the success is context and task dependent and should be defined according to the needs

plan (distance, cost, trajectory, etc).

We define four metrics, summed up in Table 2.1, allowing to measure the current task and action effectiveness. Three of them are means to measure how the progress towards the goal of a task or an action varies. Indeed, they are good indicators for the interaction quality as, when executing a task or an action, if the agents are not getting closer from the goal or even diverged from it, it means that something goes wrong. There are three different metrics because the one to use depends on the type of task or action. The fourth metric allows to compare the current execution duration to the standard execution duration of the task or action, based on durations measured during previous executions.

Metrics to assess the progress towards the goal We defined three different metrics to assess the progress towards the goal. The first one allows to assess the progress towards the goal of geometric-based actions. The second estimates the progress by using the remaining time to reach the goal. Finally, the last one measures the number of remaining steps (actions or substasks) before achieving the goal of a task.

Distance-to-Goal When an agent is performing a geometric-based action such as a movement, observing if the agent is getting closer to the target position over time provides a useful information about how well the action is going. Therefore, we introduce the *Distance-to-Goal* ΔDtG metric:

$$\begin{cases} \Delta DtG(t=0) = 0 \\ \Delta DtG(t) = \max(0, \Delta DtG(t-1) - 1) \\ \qquad\qquad \text{if } path\_length(t) < path\_length(t-1) \\ \Delta DtG(t) = \Delta DtG(t-1) + 1, \text{otherwise.} \end{cases} \tag{2.3}$$

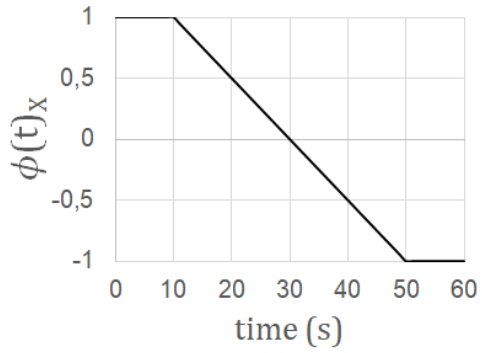
with $path\_length(t)$ the length of the path leading the goal at time t (e.g. which can be given by a reactive motion planner [Khambhaita 2020]). The metric lower bound is 0. If at time t the agent is closer to its final position than at $t-1$, i.e. progressing towards their goal, the metric is set to decrease or to remain equal to 0. Now, if the agent has not moved or is even further, the metric increases. The closer the metric value is to 0, the better it is, as it means the distance to the goal has decreased over time. We chose to not directly compute the difference between $path\_length(t)$ and $path\_length(t-1)$ as the results would be very different whether it is an action implying a long path or a short path.

Time-to-Goal This measure is intended to estimate the progress of a given task or action towards its goal based on the estimation of the remaining time to reach it. It compares the current estimated time to goal with the initial estimated time to goal taking into account the current task duration. As so, it is possible to measure the variation compared to the initial plan. We define the *Time-to-Goal*

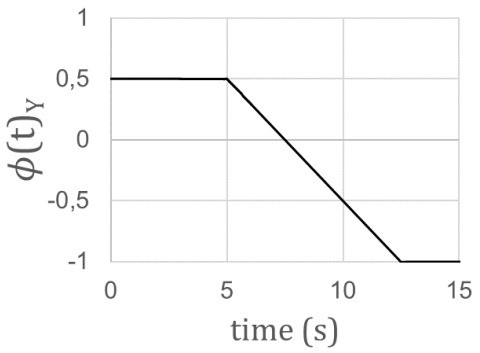
ΔTtG as:

$$\Delta TtG(t) = \max(0, e(t) + TtG(t) - TtG(T_0)) \tag{2.4}$$

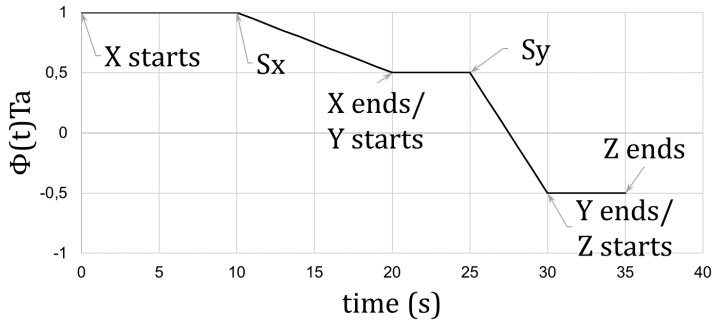
with $e(t) = t - T_0$ the task execution duration (time elapsed since the beginning of the task), $TtG(t)$ the current time to the goal, and $TtG(T_0)$ the initial planned time to goal. In our work, $TtG(t)$ and $TtG(T_0)$ are provided by a reactive motion planner [Khambhaita 2020] because we used the metric for navigation but it could be provided by other kind of planners.



(a) Plot of $\phi(t)_X$ of the subtask X lasting 60 seconds, with $SD_X = 10sec$, $V_X = 0.5$ and $\alpha = 1$



(b) Plot of $\phi(t)_Y$ of the subtask Y lasting 15 seconds, with $SD_Y = 5sec$, $V_Y = 1$ and $\alpha = 0.5$



(c) Plot of $\Phi(t)_{Ta}$ for a task composed of a sequence of three subtasks X, Y, Z: the duration of X exceeded $SD_X = 10s$ and reached $20s$, the duration of Y exceeded $SD_Y = 5s$ and reached $10s$, finally the duration of Z was less than $SD_Z = 10s$

Figure 2.3: Examples of plots of the ϕ and Φ functions

Steps-to-Goal One way to estimate the remaining distance to the goal for a task is to count the number of remaining substasks or actions (depending on the relevant scale) to perform. In addition, one can add a factor which estimates the weight (or effort needed) of each action or subtask. These weights can be determined by the designer, provided by the planner, etc. Then, the *Steps-to-Goal* $\mathcal{D}$ of a task

can be computed as time t:

$$\mathcal{D}(t) = \frac{\sum_{i=1}^{c} \mathcal{W}_i}{\sum_{i=1}^{n} \mathcal{W}_i} \tag{2.5}$$

with $\mathcal{W}_i$ the weight of a subtask/action i, c the number of completed subtasks/actions and n the total number of planned subtasks/actions.

Deviation from standard duration We introduce here a metric to measure the deviation from standard execution duration, the *Deviation from standard duration* ϕ for subtasks/actions and the *Deviation from standard duration* Φ for a whole task. This measure is intended to represent the degradation of the quality of execution of a HR task when its duration exceeds a certain time.

To each subtask/action a_i, we associate two attributes whose values are defined by the designer: a soft deadline SD_i and a decreasing quality speed V_i. If, at time t, the execution duration $e(t) = t - T_0$ of a substask or action a_i which has started at T_0 exceeds SD_i, the quality will decrease over time at speed V_i:

$$\phi(t)_i = \max\left(V_i * \frac{-\max(e(t) - SD_i, 0)}{SD_i} + \alpha, -1\right) \tag{2.6}$$

where α is the value initial value and the upper bound (as at $t = 0$, $\max(e(t) - SD_i, 0) = 0$) of ϕ_i, when the subtask/action a_i starts.

Then, we define a metric Φ for a task. It is an aggregation of the ϕ_i computed for each performed subtask/action a_i of the task. At any moment, Φ can be seen as a memory of the previous steps, so the initial value α of a_i is equal to the final value of ϕ_{i-1} of the previous subtask/action a_{i-1}, $\alpha = \phi(T_{final})_{i-1}$.

We can notice that it is not possible for this metric to increase over time since it memorizes the values of the previous actions. However, the total computed QoI can get higher thanks to the other metrics. Moreover, ϕ can be used independently of Φ. In such a case, the initial of value α of ϕ can be set to 1.

Three examples are given in Fig. 2.3. Fig. 2.3a and 2.3b represent $\phi(t)_X$ and $\phi(t)_Y$ for two independent subtasks X and Y. Fig. 2.3c is a plot of $\Phi(t)_{Ta}$ for the task Ta composed of the subtasks X, Y, Z with $SD_X = 10s$, $V_X = 0.5$, $SD_Y = 5s$, $V_Y = 1$, $SD_Z = 10s$ and $V_Z = 1$.

Chapter 3

A direction-giving robot in a mall

Contents

This chapter is from an article submitted to the User Modeling and User-Adapted Interaction (UMUAI) Journal. This work has been achieved in collaboration with Guillaume Sarthou, Guilhem Buisan, Phani-Teja Singamaneni, Yoan Sallami, Kathleen Belhassein, and Jules Waldhart. In this chapter, we first give an overview of the European H2020 Project MultiModal Mall Entertainment Robot

(MuMMER)[1]. We then present the components developed by the LAAS-RIS team. They are not my contribution but all of them are in interaction with the Supervisor.

3.1 Introduction

In large scale indoor environments, like museums, shopping malls, or airports, the presence of large interactive screens, maps, or signs underline the importance of providing information on itineraries. However, orienting and reading maps to find one's own way may be challenging. As for signs, the wanted written information may not be within sight. People also look for information not available on visual media such as the location of a given product. That is where the robot has a role to play, bringing a new way to help people to get their bearings in large indoor environments such as shopping malls.

Therefore, in the context of the European H2020 Project MuMMER[2], we developed and deployed a social service robot in one of the largest malls of Finland, Ideapark in the city of Lempäälä. This social robot is able to engage, chat with people, and guide them. We will not talk about the two first mentioned behaviors, developed by our project partners, but focus in this paper on the direction-giving.

As the mall has approximately 1.2 kilometers of shopping and pedestrian streets and more than 150 shops, people get easily lost. In such a large environment, having a robot guiding customers to their wanted destination would be time-consuming for the robot and would prevent this resource to be available for as many customers as possible. Inspired by the manner in which the mall employees perform this activity, we chose the solution to have a robot not accompanying people to their desired destination but rather verbally describing the route while grounding it with pointing gestures. If necessary, it moves a few meters inside its dedicated area (Figure 3.13) to improve the perspective sharing with the human when pointing at a landmark, and therefore to improve the human understanding of the route. These features are unique to a robot and cannot be found on a map or an interactive screen. To endow the robot with such abilities, we built a complete implementation of a robotic architecture that has been deployed in a real-world environment, the Finnish mall. There, it ran for three months, three days a week.

All along the process, we elaborated and built the system based on the main principles and ingredients which have been identified and are investigated by the Human-Human Joint Action community. We also conducted preliminary studies and used the Joint Action perspective to analyze how human guides would achieve such an activity at the place where the robot was intended to be deployed. This was possible essentially because we were able to combine the results of the JointAction4HRI[3] project with the MuMMER project.

to remove once written in chapter 1

[1] `http://mummer-project.eu/`

[2] http://mummer-project.eu/

[3] It is a multi-disciplinary project which gathers philosophers, developmental psychologists and roboticists. `https://jointaction4hri.laas.fr/`

Our claim is that such an approach is relevant in the way the joint action principles provide pertinent guidelines and it is possible to effectively elaborate models and implement systems based on them. The output is a complete robot architecture that integrates a number of components implementing the main decisions and behaviors which have been identified. Each of them makes use of various models and decisional algorithms, all integrating explicitly human models and joint action principles and mechanisms.

The chapter is constructed as follows. In Section 3.2 we provide background information about robot guides and direction-giving task and discuss about how the human partner has been considered. In Section 3.3 we discuss how we model the direction-giving task as a human-robot joint action. We analyse the task based on human-human exploratory studies and decompose it into a succession of precise subtasks in Section3.4. An overview of the resulting architecture and a description of its components are presented in Section 3.5. Then, we present in Section 3.6.2, the integration of the overall architecture into a physical robot and the steps until its final deployment "into the wild". In Section 3.7, we show how we used this task to implement the QoI Evaluator presented in Chapter 2. Finally, we present the user study we performed with 35 participants and its results.

1. March 2018: beginning of the design and implementation of the direction-giving task

2. September 2018: First tests of the task on the field, i.e., in a Finnish mall

3. June 2019 and September 2019: New tests of the direction-giving task on the field

4. From September to December 2019 (project formal end): The robot autonomously ran three days a week in the mall (with only remote monitoring of the robot performance by our team for debugging and tuning)

   (a) November 2019: Integration in the *Supervisor* of a preliminary version of Quality of Interaction Evaluator implementing the model described in Chapter 2
   $\Longrightarrow$ version 1 of the QoI Evaluator

   (b) From November 2019 to December 2019: Around 350 direction-giving tasks were performed with usual mall customers. Bug corrections and tuning of the direction-giving task. This allowed us to improve the QoI Evaluator thanks to: (i) data collection of task failures and standard durations of the subtasks executions (ii) lessons drawn about metric definitions and choices.
   $\Longrightarrow$ version 2 of the QoI Evaluator

5. January 2020: User study with 35 participants to compare three direction-giving task robot behaviors, allowing to log interactions at the same time we

could monitor them[4]. End of the project.

6. March 2020: Refinement of the QoI Evaluator, i.e., improvement of the metric functions and tuning of their parameters. In the lab, with the same direction-giving task than the one used in the mall, comparison of the QoI computed by the robot when it is dealing with an "ideal" human, a "confused" human and a "non-compliant" human.
   $\Longrightarrow$ version 3 of the QoI Evaluator

3.2 Related work

A number of contributions have proposed robot guides, from the first museum guides [Burgard 1999, Thrun 1999, Siegwart 2003, Clodic 2006] to more recent robot guides in large areas [Bauer 2009, Triebel 2016]. For example, [Chen 2017] presented a guiding robot in a shopping mall where it accompanied the customer to the desired location and pointed at the shop. Another example is a shopping robot helping people to find products among the aisles of a store [Gross 2009]. However, the focus in these contributions is mainly the fact that the robot is challenged to navigate until the goal destination with the presence of humans. Efficient mapping and localisation in large areas, social navigation are the main concerns. This is different from our needs where the robot is voluntarily constrained for its motion to a limited area with a focus on conveying to the human the pertinent information to reach by herself the desired place.

Direction-giving tasks have been investigated in the human-robot interaction community. [Kopp 2007] describes an embodied conversational agent giving route directions using deictic gestures. A number of key contributions have been developed over the years by ATR-IRC within the Robovie robot and project. First, [Okuno 2009] developed a model for a robot providing route directions, integrating utterances, gestures, and timing. The experiments explored the influence of gestures and highlighted the importance of timing in the directions-giving task. Then, [Kanda 2009, Kanda 2010] implemented a guiding behavior as part of a wider system with the robot pointing toward the first direction to take and saying "please go that way" and then, continuing its explanation by saying "After that, you will see the shop on your right.". Their robot also gave recommendations for restaurants and shops based on customer tastes. In their following work, [Morales 2011] presented a route perspective model attempting to represent humans' concept of route and visibility of landmarks, which they believed to match people's perception of the environment. Then, [Matsumoto 2012] developed a robot able to follow a user while inferring their memory recall of shops in the visited route. When the user asked the location of other shops, it gave the route description with references to the known locations inferred with the model of the user's memory recall. Finally, [Satake 2015b] showed a complete architecture of an information-providing robot

[4]The QoI Evaluator was running in background, it was not the purpose of the study.

able to move around a square in a mall composed of: a map, an ontology, a speech recognition system (operated), a dialog manager, a localization module, and a people tracker. As in their previous works, the robot verbalized utterances and used deictic gestures to give route directions.

Let us also mention [Bohus 2014], a robot providing verbal directions to people using deictic gestures coupled with spoken references. For example, the robot said "Go to the end of this hallway", executing a pointing gesture at the same time, and then continued the explanation with sentences such as "Turn right and keep walking down the hallway". [Iocchi 2015] mentioned both guiding and direction providing as use cases of their system.

Numerous other contributions can be found but, only a few of them propose full architectures for an autonomous direction-providing robot, the most complete one being the Robovie robot presented above.

Still, to the best of our knowledge, no system tackles the overall guiding-task with flexibility. Indeed we claim that it is important for the robot to reason about the current and desired perspectives of the human and the robot and to be able to pro-actively propose to the human a pertinent placement. This is one of the basic bricks of our system and it is strongly linked to the key principles of Joint Action which involve the ability to establish and monitor joint attention, and to conduct a multi-step task achievement involving contributions of both agents. Besides, it is the duty of the robot to permanently adapt to human needs and preferences and to synthesize acceptable behaviours.

3.3 Rationale

We briefly summarize here the main issues involved in Joint Action which we exposed with more details in 1.

The design of our system has taken into account the results of several user studies involving human guides of the mall (see Section 3.4). Indeed, it could be of interest to have a robot performing in the same way as a human guide does. "If robots could display predictable behaviours that are in line with human's expectations based on their models of human joint action, the resulting interaction would achieve greater naturalness" [Curioni 2019] and "human agents would then be able to apply predictive and adaptive processes acquired in human interactions to the interaction with robots" [Curioni 2019]. Here, we also take advantage of the fact that the robot is a humanoid and the human anthropomorphizes the robot behaviour (whatever we do). However, it is not always possible or desirable for a robot to imitate what a human would do at its place. It could let people think that the robot has more capabilities than it really has. In that way, besides the imitation, it could be desirable for the robot to exhibit its limitations, e.g. saying that it is able to provide you direction into the mall (and nothing else).

Participating agents in a joint action need to represent not only what they will do but also what will be performed by the others. Doing so, they also need to be able

to consider the combined effects of their respective actions [Pacherie 2012]. Joint action involves representations of the other agents who are actually and potentially involved. Shared task representations provide control structures that allow agents to flexibly engage in joint action [Knoblich 2011].

In our task, the robot has a role, it is a guide and the human is a customer with a need to find a direction. The joint action is not symmetric, there is a difference of knowledge and skills between the two agents. [Curioni 2019] raises the fact that "task asymmetry is an important factor to consider when investigating complex joint action settings because it drives the systemic emergence of communication and coordination dynamics (for example in the form of task distribution)". At the supervision level (see Section 3.5.7), we modelled which part of the task falls to the robot and which part of the task falls to the human. We could also infer that knowing the robot role as guide, the human would be able to infer what it is entitled to do. This way, we could consider that they share the route description task representation. Another important point is that "shared task representations not only specify in advance the individual parts each agent is going to perform but they also govern monitoring and prediction processes that enable interpersonal coordination in real time." [Knoblich 2011]. Our system handles that monitoring and prediction in its supervision component(see Section 3.5.7).

Joint attention provides a mechanism to form shared perceptual representations of the situation. "The phenomenon of joint attention involves more than just two people attending to the same object or event. At least two additional conditions must be obtained. First, there must be some causal connection between the two subjects' acts of attending (causal coordination). Second, each subject must be aware, in some sense, of the object as an object that is present to both; in other words, the fact that both are attending to the same object or event should be open or mutually manifest (mutual manifestness)" [Pacherie 2012, p. 355].

As explained above, joint attention comes with two requirements: causal coordination and mutual manifestness. We can consider that the engagement in the interaction session represents the causal coordination. Then, at least on the robot side, we could argue for mutual manifestness. Indeed, as we will explain, we give the robot perspective-taking abilities, and abilities to find out where and how the human and itself should be placed to share a joint attention relative to a landmark. With those requirements fulfilled, "by attending jointly, co-actors establish perceptual common-ground and become aware of each other's action opportunities and constraints" [Curioni 2019] (they also use the expression "spatio-temporal common ground"). Joint attention and in fact in our case, the overall interaction process by itself (aka its unfolding), could be seen as an continuous process sustaining under-construction common-ground.

In our system, the situation assessment component provides visuospatial perspective-taking. It computes, from the robot point of view, a number of facts regarding what the robot is looking at, which landmark is visible to it, what is present at its proximity, etc. It also computes the same information from the perspective of the person interacting with it. This way the robot is able to infer, based on its

Figure 3.1: Picture from the second Human-Human study [Belhassein 2017]. Here, the guide is giving the route description to reach a given shop by pointing at it. Positions regarding the target and the customer, as well as gazes and pointing gestures, were analyzed.

own models, which information is shared (or not) with the person it interacts with.

3.4 Designing direction-giving behavior in a shopping mall

3.4.1 What we learnt from humans

In order to inform the design and implementation of the pertinent functions and their articulation, two human-human exploratory studies were conducted in collaboration with VTT Technical Research Centre of Finland. It allowed us, in addition to the study of the existing literature, to enrich our knowledge on effective route descriptions and how they can be used in the very context of the actual robot deployment environment.

The first pilot study consisted in a human guide providing route information. It was carried out close to the future location of the robot in order to avoid biases linked to the location or the environment. Based on preliminary interviews with guides working at Ideapark, a list of 15 shops often requested by customers was selected. The preliminary experiment consisted of one participant asking for shop directions to a guide working at the mall information booth. Two researchers, as participants, and two guides took part in the experiment. The two guides were instructed to give guidance as they would normally do. The situations were video recorded and the guides were briefly interviewed after the sessions. The video

analysis focused on non-verbal communication, and in particular the different types of gestures used to give guidance, the positions of the two protagonists in relation to the target shop and their interlocutor, and the gazes alternation. [Belhassein 2017] gave the first indications to consider for the robot guidance to be effective and understood by customers, resulting from this pilot study. For example, this pilot study allowed us to notice a preferential use of the ipsilateral hand to the visual field of the target. In line with the existing literature on gestures studies, we also noticed that deictic gestures were naturally more frequent than iconic gestures or beats, while metaphorical gestures were rare. As shown by [Allen 2003], the hand used to point a referent was oriented vertically in the case of stores (vertical referents) or directed actions such as a path to take or turns, whereas in the case of horizontal referents (e.g. escalators), the hand was oriented horizontally (palm facing the ground).

A second exploratory study was then carried out adding complex situations (e.g. two customers requesting directions at the same time, two different shops in the same request, or someone who interrupts the conversation between the guide and the customer). Again, social signals were analyzed (see Figure 3.1 for an example). The protocol used and the results have been published [Heikkilä 2018, Heikkilä 2019]. By analyzing the sequencing of the whole interaction, this second study showed the guide pointing the general location of the target first, before explaining and pointing the different stages of the path to take to get there. Then, the sequencing of the route description itself showed that a first deictic gesture on a visible passage (corridor, or if the shop requested is on the second floor, the escalator) preceded the explanations about the directions to take. The most interesting results concerned situations of confusion and misunderstandings. Indeed, several elements might be sources of confusion for the customer, such as using only one transmission channel (e.g. gesture without speech), the choice of landmarks which are not always appropriate, if there are several route descriptions in the same explanation, or when the distance is not specified.

3.4.2 Design of the collaborative task for a direction-giving robot

From the analysis of human-human direction-giving and through an iterative design process, we designed and implemented our directions providing robot. Our model of the collaborative task can be represented as a succession of subtasks, as shown in Figure 3.2. This figure also exhibits the incremental refinement of the task into a sequence of human-robot interactive actions. The aforementioned subtasks are:

1. **Establishing the shared goal**: In this first step, the human and the robot negotiate and establish a shared goal. Specifically, the robot tries to determine precisely the place – we called it the *target* – it should give directions for. This is immediately completed if the human directly asked for a known shop. Several verbal exchanges can be necessary in case the person asked for a kind of shop (e.g. restaurant) or a product or in case the robot has not properly understood the name of the place and needs to disambiguate.

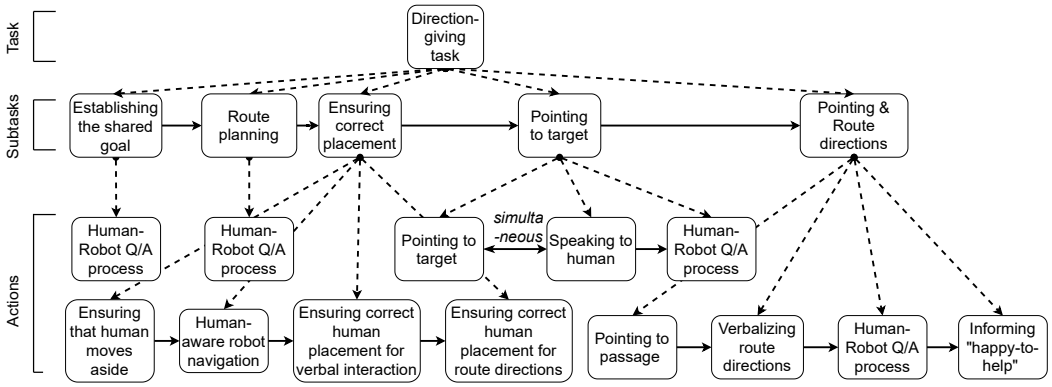


Figure 3.2: The representation of the direction-giving task as a hierarchical task network with task, substasks and actions levels. All the horizontal arrows are sequential links and the rest are decomposition ones.

2. **Route planning according to the human willingness and ability to climb stairs**: As the robot role is to help people, adapting to them, it needs to ensure that they have the abilities to follow the route it will indicate to them. So, first the robot computes the best route to the target and then checks the presence of stairs in it. In case there are, the robot enquires whether the human can or want to climb them or not. If they cannot or do not want to, the robot computes a new route without any stairs. The planned route contains a first *passage* (i.e., a corridor, a door or an escalator) which the robot will try to point.

3. **Ensuring correct placement**: The second human-human study, mentioned in Section 3.4.1, highlighted the fact that human guides point to a visible *passage* before giving the route directions. Thus, we endowed the robot with this ability as described further in the item 5 of this list. In order to be in good conditions while performing this item 5, that is to say to ease the human understanding of the directions, the robot seeks better positions for the human and itself. It does so by computing a position for the human, considering their visual perspective of the passage. The robot computes a new position for itself as well, to form a triangle whose vertices are the planned robot position, the planned human position and the passage, as shown in Figure 3.1 and Figure 3.3. After having computed these positions, the robot moves, and as they both are engaged in the task, expects the human to join it once its position is reached; it calls them if they do not. As the human might not be at the exact position computed for them, the robot checks their visibility of the passage. In case their visibility is too low, the robot will adjust their position thanks to verbal instructions (i.e., come closer, move back). Figure 3.3 illustrates the initial and final positions of both agents, in a lab context.

4. **Pointing to target**: Following the sequencing obtained from the aforemen-

tioned human-human study, the robot first points in the target direction, along with a brief sentence. As the robot is a helper and it is involved in a joint action with the human, it needs to ensure that its actions produce their expected results. In this case, if the robot computed that the target should be visible from their position, it checks that the human has seen it, either by monitoring their perspective or by asking. In case of a negative answer, it will point again.

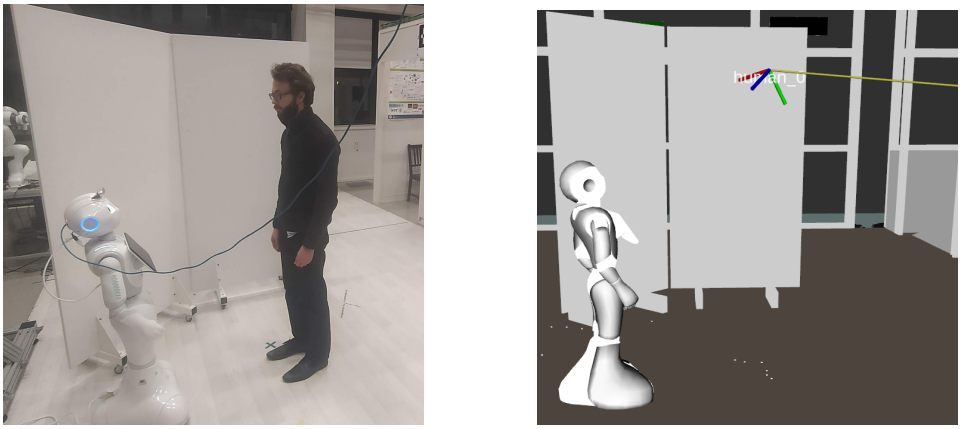
5. **Pointing to passage and giving route directions**: Still following the sequencing from the study, when the target is not in the same physical space as them, meaning that there is a passage on the way to the target, the robot points to this passage and then verbalizes the route directions. These directions take into account the orientation the human will have and describe the route (e.g., take the corridor on the left side). Finally, the way they are built (i.e., the order of the steps, the keywords to use...) is also based on the human-human study. Here again, the robot ensures that the route directions have been understood by asking the person about it or if the passage has been seen if there is one. In case of a negative answer, it will point and give the route directions again. Finally, the robot ends the task with a "happy-to-help" short sentence.

To endow a robot with the abilities described above, to build a robotic architecture embedding all these aspects, is a challenge. We tackled it with the architecture presented in the next section.

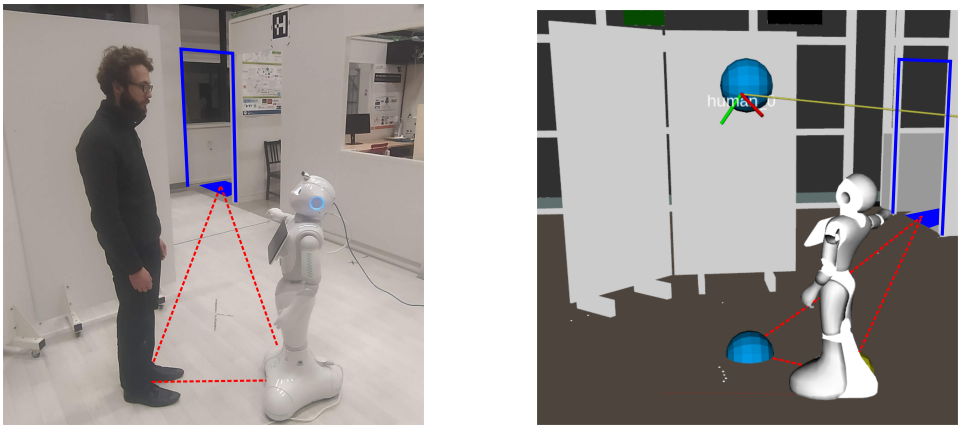
3.5 The deliberative architecture

In this section, we present the robotic architecture developed to handle the direction-giving task. This architecture relates to Beliefs, Desires, Intentions (BDI) architectures. As explain by [Wooldridge 1999], such a kind of architecture is primarily focused on practical reasoning, meaning the process of deciding step by step which action to perform to reach a goal.

The figure 3.4 represents the architecture, its components, and their interconnections. Communication between components relies on ROS. In this chapter, we only present the components developed by the LAAS-RIS team, represented by the colored blocks on the architecture. First, we present the two knowledge representations in the form of geometric and semantic representations. Next, we introduce the components related to the sensorimotor layer. It is the situation assessment and the physical resource manager. Then, we present the components related to the deliberative layer. They are the Human-Aware Navigation, the SVP (Shared Visual Perspective) planner, the Route Handler, and among the key components, we finish with the supervision and control system, designed to operate human-robot joint tasks in a joint action context.



(a) Initial positions of the human and the robot. The human asked the robot for route directions to a target behind him.



(b) The robot and the human are in their final positions. The blue spheres are the computed position for the human by the robot. The robot is pointing to the passage (in blue frame). We can observe the triangle formed between the human, the robot and the passage (the blue area on the floor) as in Figure 3.1 where two humans are in a triangle formation.

Figure 3.3: Initial and final positions of a direction-giving task in the lab context. On the left are pictures and on the right screenshots of Rviz[5](a 3D visualization tool for ROS.)

3.5.1 Environment representation

For a service robot providing directions to people, we need information to understand humans' need, information to compute the route to the goal, and information to compute the visibility of both agents to plan the pointing position. To understand the needs of a human wanted to be guided, we need information about the type of stores and the sold items. To provide so, [Satake 2015a, Satake 2015b] used an ontology. To compute the route to the final destination, [Matsumoto 2012] or [Okuno 2009] used a topological map. Each node of the graph is related to a 2D position of the environment. To estimate the human visibility of elements anywhere

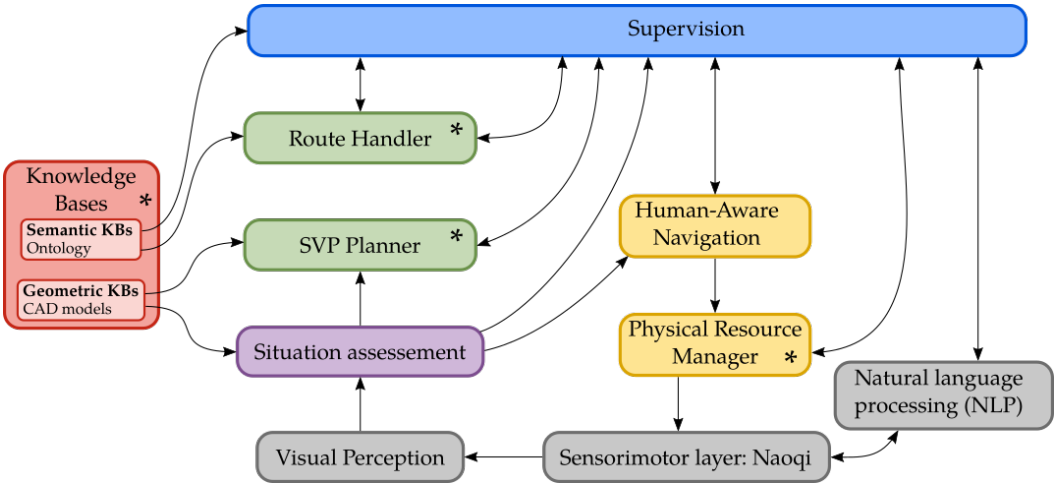


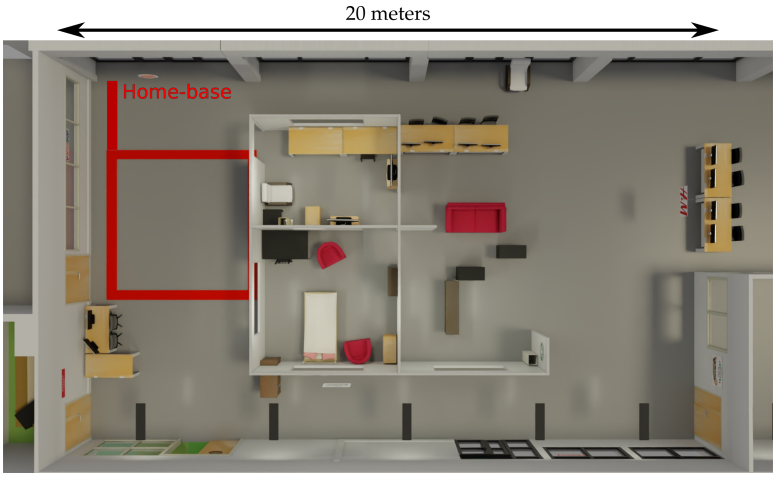
Figure 3.4: The general architecture developed for the robot guide. The components presented in this chapter are the colored blocks. The red components with the symbol * are the ones on which I participate. The visual perception and dialogue components have been respectively developed by IDIAP and HWU and are described by [Foster 2019b]. Naoqi is a Softbank Robotics software.

in the environment, [Matsumoto 2012] used a simplified 3D model where shops are represented by 3D polygons. In our implementation, we only used two types of representation of the environment: a **geometric** and a **semantic**.

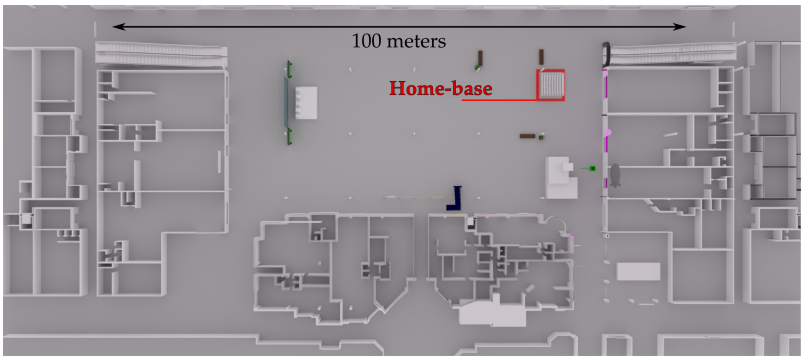
Since the final deployment of the robot was in a Finland mall, we have built an mockup mall in our lab for development purposes. By mockup, we mean that shops signs have been displayed in the laboratory to create configuration similar to the real mall. The representations describe hereafter have thus been created both for the real mall and the mockup one.

3.5.1.1 Geometric representation

The geometric representation is used to compute the visibility of elements of the environment from different positions needed for the pointing of landmarks. However, because the robot does not accompany the person to the final destination and therefore does not move much, the possible visibility of the two agents is limited to their immediate environment. For this reason and due to the large scale of the Finland mall, we chose to geometrically describe only the subpart of the global environment that could be visible from the interaction area. For the rest of the environment, we represented the shops with 3D points only. These points are enough to point in the right direction. The resulting geometrical representation is a three-dimensional mesh model, as shown in figure. 3.5a for the mockup mall and in figure 3.5b for the real one. We have represented in the 3D model all the elements that could hinder visibility, such as poles or panels. In this way, we can precisely emulate human visibility. The model was created from the architectural plans first and then refined with measurements in the mall.



(a) The 3D mesh model of the mockup mall at laboratory. The red square represent the interaction area as a square of 4 meters per 4 meters. Signs representing the shops have been place all around the environment.



(b) The 3D mesh model of the real mall in Finland. The entire mall having a size of 528.6 meters per 247.5 meters on two levels, we have only modelled the part which can be visible from the interaction area. It results in a model of 150 meters per 69 meters.

Figure 3.5: We have built a mockup of the Finnish mall environment in our lab in order to be able to test and debug the direction-giving task in our lab. This environment comprises a two-level area with corridors, "shops" , passages, stairs, open central space and consequently allowed us to run realistic guiding scenarios.

In order for the pointing planner to compute the visibility of the landmarks used for the route description, stairs, escalators, elevators, and store signs are represented each by a single mesh while the rest of the building is a unique 3D mesh. This means that a store is said to be visible if we can see its sign, which we think to be the most relevant element to see to recognize a shop.

The 3D model is also used to generate a navigation map, constraining the robot to move in the interaction area while avoiding obstacles in it.

3.5.1.2 Semantic representation

As [Satake 2015a], our semantic representation is based on ontology. An ontology allows to define classes representing general concepts (e.g. Restaurant), individuals/entities being classes instantiations (e.g. Burger_King), and properties linking two entities (e.g. Burger_King isIn Ideapark). To provide storage and an efficient way to manipulate the ontology and reason about it, a lightweight software has been developed, called Ontologenius, presented by [Sarthou 2019b]. It makes it possible to share the semantic knowledge among all the components of the architecture, here especially the route handler and the supervision, thus enabling a unique repository of knowledge.

The ontology is first used to represent information about the stores. It allows to define and refine the shared goal of the task by understanding the client's wanted destination. Thus, the stores' types, their names, and the items they sell have been represented in it with a rich semantic. It allows for example to represent that both soda and hamburgers are sold in fast-foods, which are types of restaurants, but that soda can also be found in a supermarket. Thanks to Ontologenius, the names of concepts are defined in different languages and with synonyms for these names. It allows the robot to adapt itself to the human partner language. Moreover, Ontologenius endows the robot with the ability to recognize a set of names in natural language but that it will be prevented to use (e.g. the robot can understand a reference to "bank" when a human says it but only refers to it as "ATM" or "cash machine" since there was no bank office in the mall). In addition, this software offers a fuzzy match service based on Levenshtein distance, to help the supervision system to handle ambiguities coming from the speech to text component (e.g. it can match the word "Juwelsport" with "Juvesport"). This set of functionalities around the concepts' names facilitates the understanding of the partner's need and thus helps at increasing the quality of interaction.

In an effort to unify representations because representing as a 3D mesh the entire mall would be a complex task, we chose not to use the geometrical representation or a topological map to compute the route to the final goal but rather the semantic representation given by the ontology.

To include topological information into the semantic representation, the Semantic Spatial Representation (SSR) has been designed, presented by [Sarthou 2019a]. With the SSR, the overall knowledge is represented in an ontology with three upper classes which are: **region** (i.e. a two-dimensional area that is a subset of the overall environment), **path** (i.e. a one-dimensional element along which it is possible to move and which has a direction) and **place** (i.e. a point of zero dimension that can represent a physical or symbolic element). The **place** class has three subclasses: **path intersection** (i.e. the connection between only two paths and thus a waypoint to go from one path to another), **passage** (i.e. the connection between only two regions and thus a waypoint to move from one region to another like a door, a staircase or a passage), and **shops**. A representation of these classes is visible in Figure 3.6.

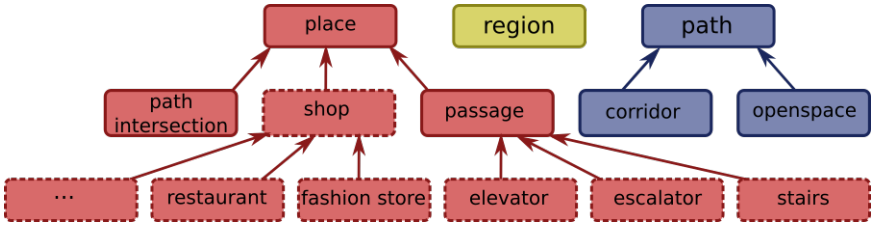


Figure 3.6: Classes for a representation of the topology of an indoor environment in a semantic description. The classes with the solid outline are the minimum classes defined by the SSR. The classes with the dotted outline are an extension of this minimal set.

An example of the final semantic knowledge represented in the ontology for a given shop is presented in Figure 3.7. We find here the identifier of the shop, the category to which each store belongs (e.g. restaurant or hairdresser), the items sold for which people ask the most (e.g. shoes or coat), and the names and synonyms in natural language and that for different languages. Moreover, thanks to the SSR we can produce the best route (in term of complexity) as well as verbalize it using a route perspective.

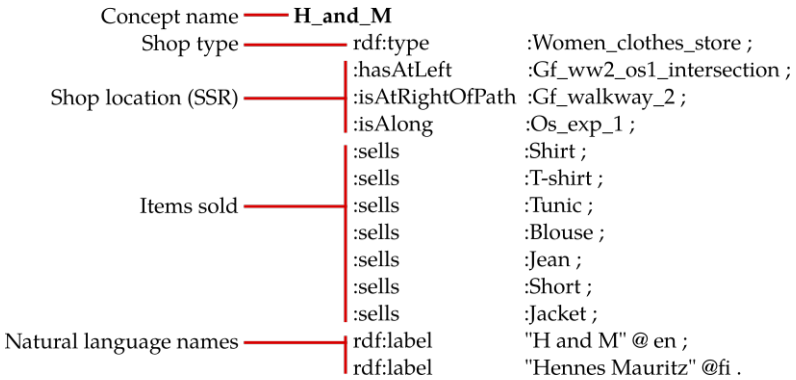


Figure 3.7: Properties for a representation of the topology of an indoor environment in a semantic description.

3.5.2 Perceiving the partner

The situation assessment component is based on the Underworld framework [Lemaignan 2018]. It aims at gathering perception information in the form of 3D position and orientation of human faces, with the 3D model and the robot state. With this information, it is able to generate the symbolics facts listed in table 3.1.

3.5.3 Managing the robot's resources

A humanoid robot such as Pepper can be seen as a composition of multiple physical components that can act independently of each other. For the pointing task, we

Predicate	Description
isPerceiving	The robot is perceiving a human
isCloseTo	The human is within a distance of 0 to 1 meter of the robot
isLookingAt	The human is looking at the robot
isInArea	The human is in the interaction area
isEngagingWith	The human is close to the robot and is looking at it

Table 3.1: Facts computed and monitored during the direction-giving task.

identified four resources: the head, both arms, and the base. At the beginning of the interaction, for example, the head is used to find people to interact with, but later it will be used to track the human with the gaze. Several components could access this resource to perform these actions. However, they do not have a global picture of the ongoing task. In this case, a resource could be used by several components at the time. Consequently, it could lead to task failures.

Moreover, in some cases, several resources have to be used simultaneously to perform a high-level action. To point to a landmark, one arm is selected to point while the other has to be lowered. The base is then rotated if the arm reaches the joint limit to point a target on its back. If at least one of the involved resources is simultaneously used to perform another action, the overall high-level action will fail as the global posture will no more be clear. For example, if the human gets too close to the robot and a component tries to move away from a little, the arm would no more point in the right direction.

Thus, the correct handling of all the resources is critical for performing the task, but it can be cumbersome for a deliberative component, such as the Supervision, to do all the micro-management required. To tackle this issue, a physical resource management system has been designed by Guillaume Sarthou and Guilhem Buisan. For each of the identified resources is instantiated a component called **Resource Manager**, having two types of input: permanent commands that can be preempted at any time (e.g. look at the head of the human interacting with it) and finite state machines which are not preemptable (e.g. set of commands to point). A component called **Resource Synchronizer** deals with actions requiring multiple resources such as the human-aware navigation which uses the head and the base. The synchronizer also reports the status of the ongoing coordination signal to the Supervisor to monitor the progress of the action. Finally, a priority scheme has been implemented to handle multiple active inputs at the same time for one resource.

The global resource management scheme is illustrated in Figure 3.8 with four resource managers and one synchronizer.

3.5.4 Describing the route to follow

In large-scale environments such as malls, route computation can lead to combinatorial explosion. Therefore, to simplify the problem we chose to divide it in two stages and to conceive an algorithm for each one. The first one computes the exist-

ing routes from one Region of the mall to another. A region is for example a floor of the building so if the robot is at the ground floor and the final destination also, this means that the algorithm will not take into account all the elements of the other floors. Then, the second algorithm uses the Region-to-Region routes to calculate the Place-to-Place route. These algorithms are presented with more details in [Sarthou 2019a].

Region-to-Region route In the SSR, **passages** (e.g. escalators, stairs) are elements of the environment connecting two regions through the $isIn$ property. With this property and a breadth-first search algorithm, Region-to-Region route finding algorithm is able to find the routes connecting two regions using only passages. It outputs a route with the format $region - place - region - ... - region$. In the example of Figure 3.9, the final routes found by the algorithm to go from Region 1 to Region 3 are:

- $region\_1 - passage\_1 - region\_2 - passage\_2 - region\_3$
- $region\_1 - passage\_1 - region\_2 - passage\_3 - region\_3$

Place-to-Place route The Place-to-Place route search is based on the Region-level search results. It decomposes each route into sub-routes of the form $place - region - place$. In our example, the division gives five unique sub-routes:

- $start - region1 - passage\_1$

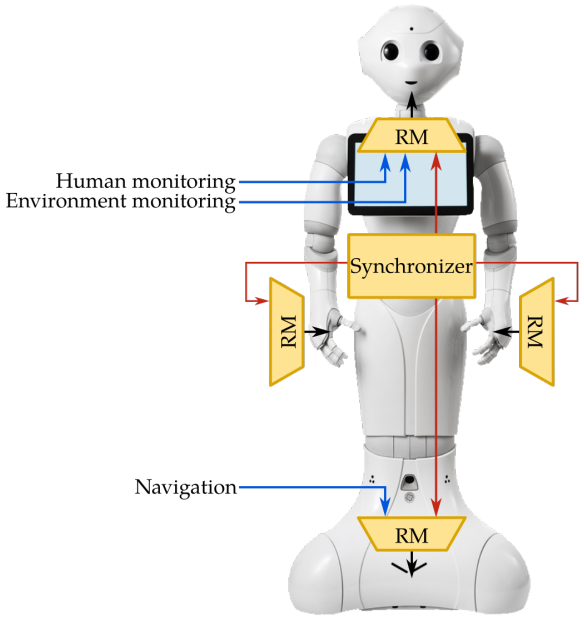


Figure 3.8: Representation of the resource management system with four resource managers and a synchronizer. The red arrows represent the state machines inputs and the blue arrows represent the inputs for permanent commands.

- $passage\_1 - region\_2 - passage\_2$
- $passage\_2 - region\_3 - end$
- $passage\_1 - region\_2 - passage\_3$
- $passage\_3 - region\_3 - end$

Then, the algorithm aims to replace each sub-route region with a succession of paths and intersections. It works on the same principle as the previous search algorithm using the $isAlong$ property instead of the $isIn$ property. Still taking the same example and focusing on $region\_1$, it can solve the sub-route $start - region\_1 - passage\_1$. $Region\_1$ is represented with its corridors and intersections in Figure 3.10. By applying the breadth-first search algorithm at the Place level, a solution of the form $place - path - place - ... - place$ is obtained. So for our example, $start - corridor\_1 - intersection\_1 - corridor\_5 - passage\_1$ is a solution for the first sub-route. By doing the same for each sub-route, we can then recompose the global routes and give a detailed set of routes from start to end.

The second place of the route – the third element of the route objects – is the one we call the passage in the description of the direction-giving task, the first salient landmark of the route to point to, which is on the way to reach the final place, i.e. $intersection\_1$ in the example.

3.5.5 Planning a shared visual perspective

When the robot has to point to a target, two criteria have to be respected. First, the human has to be able to see the target. Second, the human has to be able to look at the pointed target and at the robot without turning the head too much. It goes the same for the robot as it has to see the pointed target, meaning not to point toward a wall and be able to simultaneously point at the target and look at the human. Consequently, to point a target in its back, it has to move. The robot and the human can thus move in the interaction area during the direction-giving task, to move to a better position for pointing at the target. To find the robot and human possible positions we designed a component called the SVP (Shared

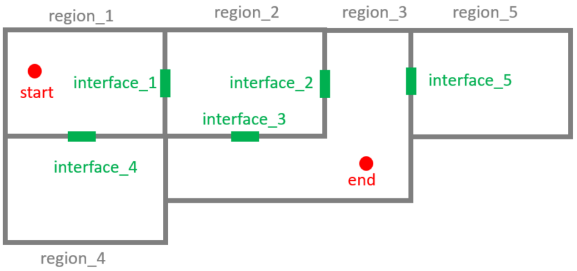


Figure 3.9: Representation of an environment at the regional level.

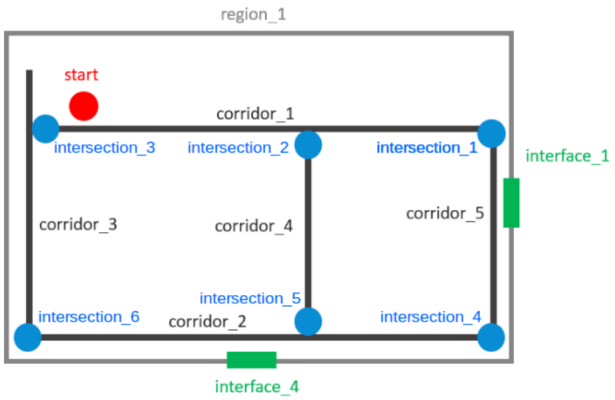


Figure 3.10: Representation of corridors and intersections in region_1

Visual Perspective) Planner, presented in [Waldhart 2019]. For the purpose of the deployment, the presented version is an adapted and slightly simplified version.

To compute the visibility of both agents, the planner has access to the geometrical representation of the environment and the agents current positions. In addition, it considers an estimated agent's maximal speed to move and a visibility threshold.

When the robot explains the route to the human and points to a landmark, they form what is called an F-formation. Kendon explains that *"An F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct and exclusive access"* [Kendon 1990]. This F-formation has been decomposed in [McNeill 2005] into two types: the social formation and the instrumental formation. While the first type corresponds to the original definition, the instrumental formation includes a physical object that all the agents can gaze at. This means that once the robot will have moved, the human will come in front of it creating a social formation in the form of a vis-a-vis (each facing the other) and when the robot will point they will change for an instrumental formation. Indeed, when both agents will reach their position computed by the planner, we want them to be able to go from one formation to the other with only a rotation; the human will not need to move again from their arriving position to see what the robot will point.

To search for better positions to reach in order to point a landmark, the planner takes three main parameters into account:

- Visibility constraint: The two agents can see either the target shop when it is the only element of the route or the passage.
- Navigation distance cost: The agents do not have to move too much.
- F-formation cost: The human-robot-target angle and a robot-human-target have to be less than 90°.

To compute the positions, the interaction area is firstly decomposed into a

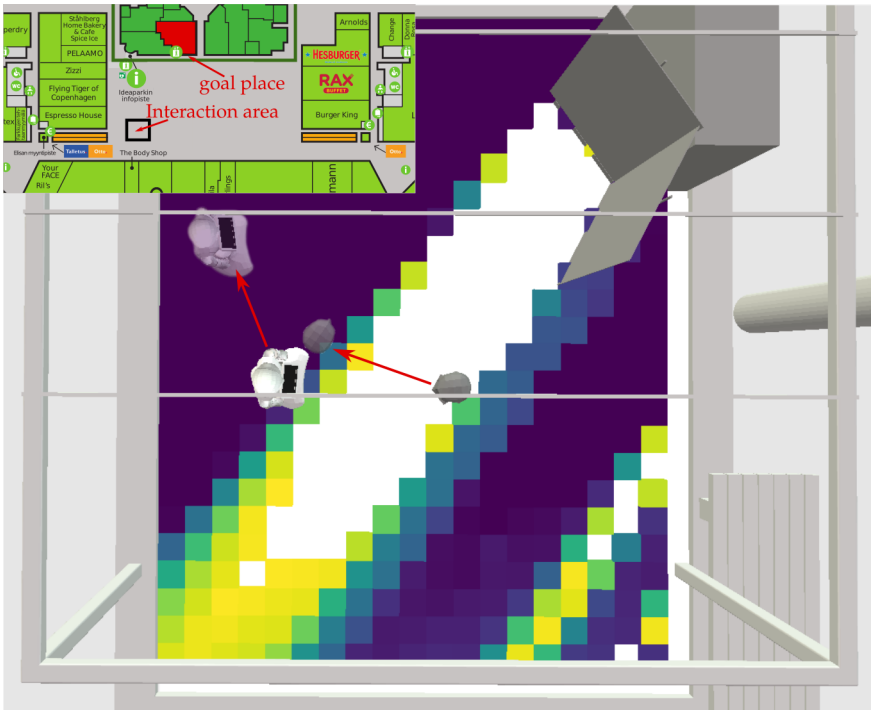


Figure 3.11: Visibility grid for a target located at the top right. The uncoloured areas represent an absence of visibility and the others represent the cost of visibility ranging from yellow for low visibility to purple for good visibility. The robot and the human in transparency on the image represent the final calculated positions while the others are the initial positions.

weighted three-dimensional (x,y for the possible positions in the area and z for the human height) grid representing the estimated human visibility of the target. The target visibility is computed offline for each position of the grid. It is based on the part that the target takes in the 360° field of view of the environment. Such grid is represented in figure 3.11 for a given human height. The white cells are positions from which the human cannot see the pointed target. The other colored cells represent the degree of visibility from the poor in yellow to the good in purple. Having the human visibility grid, the goal position is computed using a weighted cost function between good visibility and restricted distance to cross. In the example of figure 3.11, the transparent human head is the human goal position while the other is the initial position. From the initial position, the human was not able to the pointed target.

The robot position is computed in a second time, according to the human planned position. Divided the search into two steps allows reducing the search complexity. The robot position is thus constrained by the human one. It has also to respect a minimal and maximal distance to the human and minimal visibility of the target from it. Finally, the robot position is also determined regarding a cost preferring an F-formation limiting the robot reorientation, meaning that it can

point to the target keeping its torso and its chest oriented towards the human.

3.5.6 Navigate close to human

The Human-Aware Navigation component aims at moving the robot while avoiding dynamic and static obstacles in addition to proposing a socially acceptable navigation solution for the robot. For example, the robot should not pass too close to the human and should not show its back while navigating around the human. A full presentation of the planner is available in [Singamaneni 2020].

3.5.7 Robot execution control and supervision in a joint action context

subsubsectionA supervision and control system dedicated to human-robot joint tasks A service robot interacting with humans in a mall and providing directions to them needs a number of abilities to enable a smooth and efficient interaction. As explained in Section 3.3, the direction giving task is an asymmetric joint action, with the robot in the guide role and the human in the guided role. The *Supervisor*, the supervision and control system of the robot, is built taking this specificity into account, embedding a shared representation of the direction giving task. More specifically, when giving directions to a human, the robot plans its actions and the human ones and then execute its part of the plan. To be able to know if and when the human performs their actions, it monitors the action executions and interpret the information received from the Situation Assessment (see Section 3.5.2). Furthermore, in such interaction, communication is important, thus the robot communicates verbally as well as non-verbally, and listens to the human. All along the interaction, it needs to maintain a distinct mental state model for the human and itself concerning the knowledge of both agents and the state of the world. Finally, it should be able to tackle events and contingencies happening during the task and to drop it when necessary.

A *direction-giving task* occurs when the human involved in the ongoing interaction session asks for directions to a place or for locations of sold items.

voir lien avec chapter 1

voir lien avec chapter 1

3.5.7.1 Implementation of the direction-giving task and its associated actions

The Supervisor can be seen as a puppet master. It manages the robot's decisions of when to act, what to say, and what to do during the task. It decides, executes, and supervises the proceedings of the execution as well as the human actions, using and coordinating all the other components of the architecture at its disposal.

The Supervisor is implemented using Jason, so it can manipulate beliefs and reactive plans (written offline by the developer rather than planned during the interaction). For the interaction session and the direction-giving task, at execution time, plans are chosen among the ones from the plan library when triggered by an

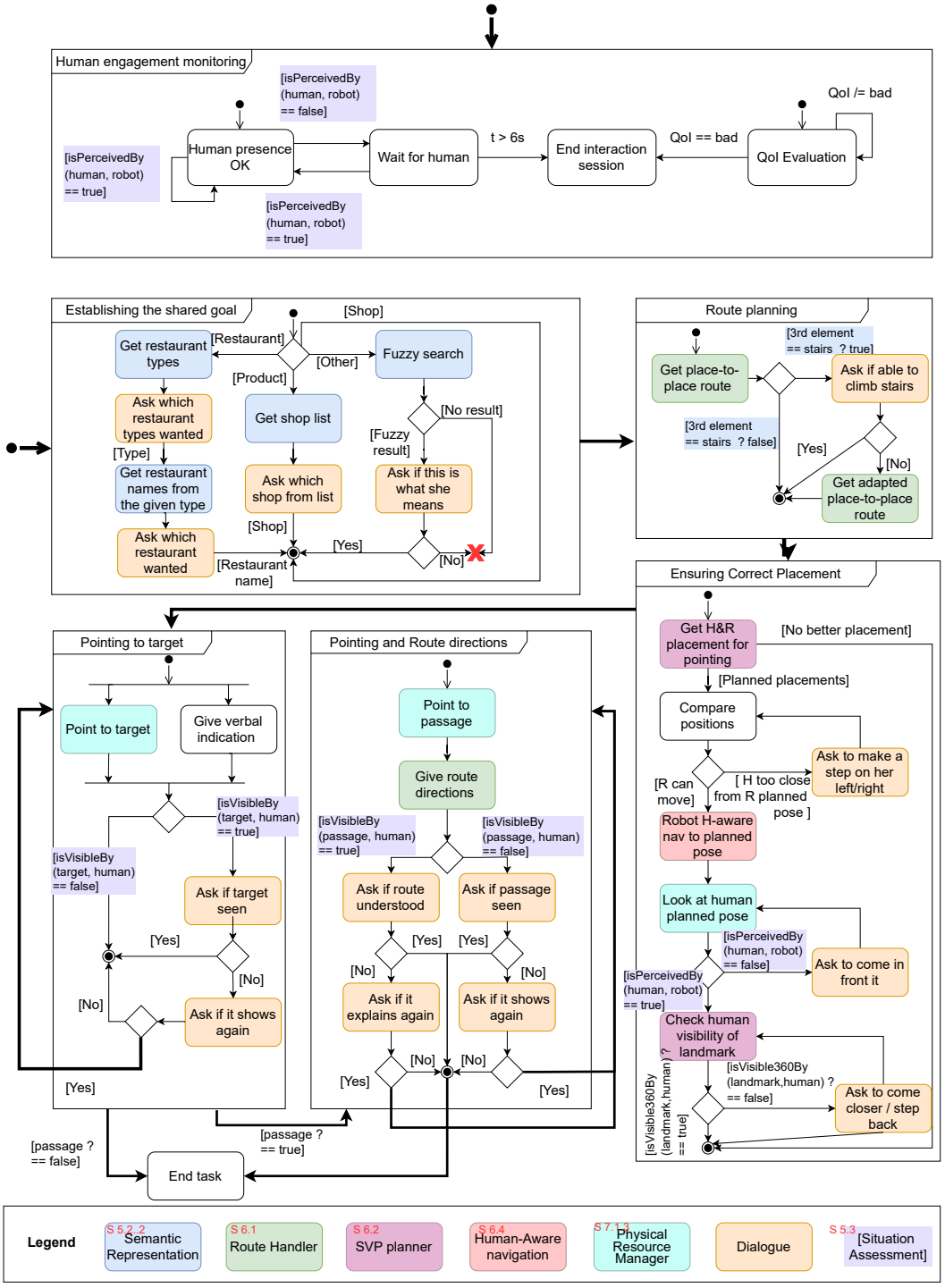


Figure 3.12: Supervisor activity diagram of the direction-giving task. Each action has a color corresponding to the component with which the Supervisor interacts to execute it. It goes through every subtasks described in Section 3.5.7.1. Also, the human engagement monitoring is represented. Texts between brackets correspond to beliefs on which depends the decision-making process. These beliefs can either be provided by other components or being the result of the Supervisor's own computations.

event or by another plan. The same plan can have multiple versions and the version to be executed is selected according to the pre-conditions (also called context). For instance, the plan *verbalization(Target)* has two different versions, one in the case where the target to point is visible and the other one in the case where it is not, and at execution time, the selected one will depend on the presence or not of the belief `visible_target(Target)` in the Supervisor belief base, as shown in Listing 3.1:

Listing 3.1: Two different plans for `verbalization(Target)`

```
+!verbalization(Target)                    // plan name
: visible_target(Target)                   // context
<- ?verba_name(Target, Name);              // belief query
say(visible_target(Name)).                 // action

+!verbalization(Target)
: not visible_target(Target)
<- ?verba_name(Target, Name);
say(not_visible_target(Name)).
```

Even though the direction-giving task is implemented with reactive plans, it can still be represented with an activity diagram, for presentation purposes. This activity diagram is visible in Figure 3.12. Each frame represents one of the steps described in Section 3.4.2. We now present their internal functioning and the interactions with the multiple components of the system the Supervisor has.

Establishing the shared goal When a person triggers a direction-giving task, they might directly ask something like "where is the pharmacy?" which allows the robot to directly establish the shared goal but, they might also ask something less precise. In the latter case, the robot needs to inquire about the human desired place to reach in order to establish the shared goal.

reminder def

When a person asks "Where is a good restaurant?", the robot presents a list of the types of food available, namely "There are casual dining restaurants, Asian restaurants, native food restaurants, hamburger restaurants, fast food restaurants, and pizzerias.". This behavior is quite similar to the recommendation behaviors of [Kanda 2009].

To be able to display this behavior, several components of the system are requested. When the Supervisor receives $\{request = restaurant\}$ as data from the Dialogue, it asks Ontologenius (see Section 3.5.1.2) for all the existing restaurant types. This list of restaurant types is sent to the Dialogue whose role is to return to the Supervisor with the type selected by the human. Finally, similarly to the way it obtained the restaurant type from the human, the Supervisor tries to get the restaurant name. Therefore, it requests from Ontologenius all the restaurants serving the given type of food. Then, this list is sent to the Dialogue whose role is to return to the Supervisor with the restaurant selected by the human among

the elements' list. It should be noted that all the restaurants of the given type are suggested to the person, even though sometimes the list is long. We thought of alternatives such as randomly giving three restaurants among the ones of the list. However, these alternatives were not allowed by the mall policy as they could not provide equality between all shops.

The same principle goes for products. For example, people can ask "Where can I buy a dress?". Then, the Supervisor gets from Ontologenius a list of shops selling dresses and passes it to the Dialogue. The Dialogue returns the name of the shop chosen by the person.

When the Supervisor receives as a goal a name it does not understand, it queries Ontologenius to try to match it to a known name as it may be not understood because of a speech recognition failure or a shortened name. For instance, thanks to the fuzzy match provided by Ontologenius, when a person asks to go to "jewelsport", the system can make the assumption that the person actually asked for "Juvesport". So the robot asks the person, "do you mean Juvesport?", to which the person can answer "yes" or "no". If yes, it starts the direction-giving task, if no it drops it and returns in chat mode.

Enquiry about human willingness and abilities to climb stairs As the robot is there to help humans, it has to adapt to their abilities and preferences such as a person with a shopping trolley will prefer to take escalators than stairs. The preferences definition is currently done through verbal communication.

To determine human preferences about stairs, the Supervisor first requests to the Route Handler (see Section 3.5.4) the possible routes to go to the target shop. The returned routes are of the form $place-path-place-...-place$. The Supervisor selects the one with the smallest cost and then checks if one of the *place* elements is stairs (i.e. the Supervisor queries Ontologenius for the element type). If it is the case, the Supervisor asks the Dialogue with finding out if the human is able to climb stairs or not. If not, it will send a new request to the Route Handler with the parameter "no stairs" and will get a new set of routes. The Supervisor selects the one with the smallest cost. This new route will have a cost equal to or higher than the first one (since it was not the route with the smallest cost in the initial request), which means the goal might be more complicated to reach or it might take more time.

Ensuring a correct placement The robot's role in this task is not only to give verbal route directions but also to point to the target and the passage (i.e. the third element of the route as explained Section 3.5.4) the person should take in order to increase the chances that they reach their destination as it helps to orientate them in space. For the pointing to be as efficient as possible, the robot computes new positions for itself and the human where the visibility of the pointed landmarks will be better (when feasible). Then its goal is to have itself and the human reaching these new positions.

In the first step of this subtask, the Supervisor requests from the Shared Visual Perspective (SVP) Planner (see Section 3.5.5) the new positions for the robot and the human, with the passage to point (or the target if no passage) and the human identifier as parameters. Then, the Supervisor compares the newly received positions with the current ones of the human and the robot – the current position of the human is provided by the Situation Assessment. In the case where the robot planned position is very close to the human's current position (< 0.5 m), the robot asks the human to step aside on the right or left, depending on the human's planned position. If the human does not move or does not go far enough from the planned robot position, the robot will ask again.

Then, the Supervisor requests the Human-Aware Navigation (see Section 3.5.6) to move the robot to its planned position. Once the Human-Aware Navigation returned that the position has been reached, the Supervisor looks for the human. It is a form of monitoring, which we show in Section 3.3 is important in a joint action. If the human is not perceived – the Supervisor did not receive from the Situation Assessment the predicate isPerceiving($robot, human_i$) – in the following seconds (6 seconds in the deployed version), the robot asks the human to come in front of it – this is the way we have chosen after several trials (other modalities like indicating to the human by a gesture where they should stand were not sufficiently successful). If the human is still not perceived after a few seconds, the robot will ask again, remaining engaged in their joint action for a while before giving up.

Once the human arrives in the robot field of view – which means that the human more or less reached their planned position since the robot is looking in the direction of it –, they might not exactly be at their planned position. In this case, their position may not be suited to properly see what the robot has to point at. To check if they are in a position good enough to see, the Supervisor asks the SVP Planner for the visibility (at 360 degrees) of the landmark to point. In the case where the SVP Planner returns that the landmark is visible, the interaction continues. Else, the robot asks the human to move forward or backward in order to adjust their placement according to their planned position. This stops when the robot computes that the position of the human will allow them to see the target. In this way, the robot tries to ensure to put the human in the best conditions as possible for the next steps, using key elements of the joint action: monitoring of the partner actions', sharing a visual perspective and showing engagement in the task.

Pointing to target As it is shown that the use of deictic gestures such as pointing improves the understanding of route directions (see Section 3.4.1), we endowed the robot with this ability.

To do so, the Supervisor requests from the Physical Resource Manager that the robot points to the target. At the same time, it generates a short sentence for the robot to say and sends it to the Dialogue. The sentence varies according to the visibility of the target such as "Here, you can see Burger King" for a visible place and "The restroom is in this direction" for a non-visible one. In this way, the robot

shares the human's perspective and takes into account the knowledge they can get from their environment in respect of the joint action principles. In this way, the human knows if they have to try to notice it from their place or take this information as an orientation indication. In order to continuously look at the human and not loose them from its sight, the robot does not turn its head towards the target when pointing.

It is important for the robot to know if it successfully communicated the information to the human. Then, it asks if the target has been seen, as it wants to ensure its action had the expected effect.

Pointing to passage and giving route directions This step is executed when there is a passage in the route returned by the Route Handler. Therefore, the Supervisor sends a route to the Route Handler which returns a verbalization of this route (e.g. "Walk through that corridor, and then, turn left. From there on, Apteekki will be on your right, straight after Glitter"). Then, as explained in the *Pointing to target* paragraph, the robot points, to the passage this time. And, at the same time, it verbalizes the route received from the Route Handler, added "in this direction" to the sentence if the passage is not visible.

As for ensuring the target has been seen, the robot wants to make sure it has been understood and leaves the possibility to the human to hear the route directions again if they need it. In the early versions, we had programmed the robot to ask if the passage had been seen and then if the route had been understood but it was too many questions that seemed useless to users. Indeed, we analyzed it as a postcompletion error [Byrne 1997], as the goal of the human was to know the route to their location, whatever actions arising after this goal has been completed are often forgotten. In the end, the first question is asked in case of a visible passage and the second one is asked in case of a non-visible one.

It may be noted in Figure 3.12 that it is possible to go in infinite loops such as Route directions - Ensuring route understood - Route directions - ... . To avoid this issue, the Supervisor prevents to return inside a step if it has already been executed a certain number of times (in the final version, 3 was the maximal number a step could be executed).

3.6 A robot in the wild

3.6.1 Pepper in Ideapark

For availability for as many customers as possible, the robot was contained in a defined place in the mall as shown in figure 3.13. A home base was designed with the participation of all the project partners. It was a 4 per 4 meters area with a 2.5m high frame structure on it. The home base included a non-reflecting carpet on the floor and an acoustic ceiling surface on the roof.

During the first deployment in the real mall, we have updated both the Geometric Representation with actual measurements and the Semantic Spatial Repre-

Figure 3.13: The pepper robot in its interaction area in the Finalnd mall, Ideapark.

sentation (SSR) by making sure the regions, interfaces, corridors and intersections were represented reflecting the actual mall topology. To ensure the correctness of the instructions given by the route handler, we generated routes from the deployment location to several shops in the mall and followed them to the destination. Inaccuracies, as well as algorithmic flaws, have been fixed using this method. We also tested the interaction in the Finnish language with our native Finnish partners and corrected some mistakes in the route verbalization.

3.6.2 The deliberative architecture embedded in a physical robot

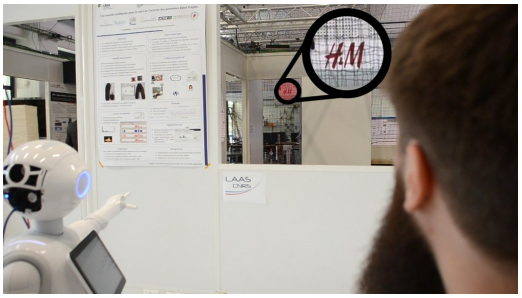
In the previous section, we presented a deliberative architecture designed to be embedded in a service robot. The purpose of this robot was to be deployed in a mall in Finland. To make this deployment successful, we did extensive tests in our laboratory where we had reproduced a part of the mall environment to be in the most realistic conditions possible[6]. Some of these emulated shop are visible in Fig. 3.14. In Sect. 3.6.2.1, we introduce the environment setup as well as the robot one. Then, in Sect. 3.6.2.2 and Sect. 3.6.2.3, we present our tests and deployment in the Finnish mall.

3.6.2.1 Environment and robot setup in the Finnish mall

Our architecture has been tested and deployed in a mall in Finland. As we explained previously, it has two abilities: chat with people and guide them, but in this paper we consider only the latter. The robot was able to interact in English and Finnish, though due to the vast linguistic differences between the two languages, the two versions have been kept separated, and the whole interaction can either be in one or the other.

[6]This setup not only was used for tests but also for public demos and even in the context of a scientific live event now accessible on `https://youtu.be/p4f3iwHht2Q?t=4495`

(a) A person being guided, the emulated shop "Zizzi" is visible in the background of the picture.



(b) A person being guided, the emulated shop "H&M" is visible.

(c) Two people simulated going to shop. The emulated shop "Burger King" is visible in the background and a small part of "Thai Papaya" is visible in the foreground.

(d) A person being guided, a sign towards the toilet and the shop "Marco Polo" are visible on the left of the picture.

Figure 3.14: Examples of emulated shops of the Finnish mall in our lab.

Hardware architecture The robot is an upgraded, custom version of the Pepper platform [Caniot 2020], which is equipped with an Intel D435 camera and an NVIDIA Jetson TX2 in addition to the traditional sensors that are found on the previous versions of the robot. We used the Robot Operating System (ROS) to enable inter-process communication between the processing nodes. All the streams (audio, video, robot states) are sent to a remote laptop which performs all the computation. The laptop has an NVIDIA RTX 2080 graphics card (for the visual perception system) and 12 CPU cores. The 4 microphone streams are processed at a frequency of 16000 Hz, and the full perception system delivers the output at 10 fps.

3.6.2.2 Pre-deployment in the Finnish mall, in-situ tests

Three integration sessions, each lasting one week have been made on site, in September 2018, June 2019 and September 2019, in the mall in Finland. The whole LAAS developer team were part of these integration weeks, along with our project partners. So, I spent around 150 hours (3 times 5 days) in the mall for software integration debugging with the other developers, and testing and debugging of the direction-giving task. During the integration weeks, only expert users (developers) interacted with the robot for testing purpose.

To have a working system in the lab and to have a working system in a real-world site are two different things. As much as a team prepare for an in-situ deployment, there will always be elements that will need to be tuned on site and unexpected bugs arising. Thus, I had to handle a lot of contingencies, diagnosing where the issue came from, repairing if it was originating from my software, communicating with the person responsible for the component having a bug if it was not from mine, and testing again.

The first step to perform for us once on site was to update both the Geometric Representation (see Sect. 3.5.1) which was previously based on architectural plans and refined with actual measurements and the Semantic Spatial Representation (see Sect. 3.5.1) by making sure the regions, interfaces, corridors and intersections were represented reflecting the actual mall topology.

Finally, to ensure the correctness of the instructions given by the route handler, we generated routes from the deployment location to random shops in the mall, and followed them to the destination. SSR inaccuracies as well as algorithmic flaws have been fixed using this method. We also tested the interaction in Finnish language with our native Finnish partners and corrected some mistakes in the route verbalization.

Component integration problematic Even though components were integrated together before getting on site, code modifications as mentioned above and intense testing can make new issues appear. So, it was essential to test the integration between all the components after this.

Finally, once everything was running quite nicely, some time has been dedicated

to fine-tune the direction-giving task, ensuring all the components could withstand running for several hours in a row, with naive users possibly interrupting the task at any stage.

3.6.2.3 "In the wild" deployment

The robot was then installed for a long-term 14 weeks deployment from September 2019 to December 2019. During this period, the robot interacted with everyday clients of the mall, who may never had the chance to interact with a robot before. The robot was active for 3 hours per day, three days a week. As it was a project with multiple partners, it was not always possible to have our direction-giving task running. The direction giving task has been available on the robot 32 days out of the 42. The days during which it was not running, the partners' software executing on the robot where the visual perception and the dialogue that we mentioned previously, and a social signal processing feature [Foster 2019a].

Nowadays, having an autonomous robot in the wild is a challenge. At first glance, we could think that if the robot is able to run smoothly for a few hours, the challenge would be met. However, there are a lot of other elements to take into account. First, how to guarantee the safety of children and elderly? How to ensure that the robot will not fell on or bump into them despite the robot sensors, hurting them? Furthermore, not only people safety is important but making sure that the robot is not damaged by people as well. People might indeed be brutal towards the robot, on purpose or not.

To tackle the "obvious" issue, making sure that the robot continuously running, it was remotely watched by an on-call developer of the project team. At the beginning of the time slot, they launched all the software on the robot. Then, they checked through component monitoring, time to time, if everything was running properly, and they were in contact with the robot guard who told them if she noticed something wrong with the robot. They also had access to a video feed of the robot home-base if needed. Thus, all along this long-term deployment, we adjusted parameters and fixed bugs, with the help of the on-site team VTT and the robot guard that tested the direction-giving task when we asked her. The bugs we encountered concerned mainly Finnish translation issues (e.g. "just after Arnold's" was translated "paikan päälle Arnolds" in Finnish but the correct way to say it in Finnish was "paikan Arnolds jälkeen" thus we changed the English sentence into "right after the place Arnolds" to be able to get this translation), shop names issues (e.g. Finnish people use the utterance "Hennes Mauritz" and not "H&M" which was the name in the robot ontology originally) and route issues (e.g. a route has one more turn than it should have).

The project consortium tackled the two safety issues (people safety and robot safety) by hiring a "robot guard" and by putting a sign notifying parents to not leave their children alone with the robot. During the robot active hours, this guard employee was physically present to ensure people were respectful towards the robot, i.e. not hitting it or pulling it, to watch the kids who may get too close to the robot

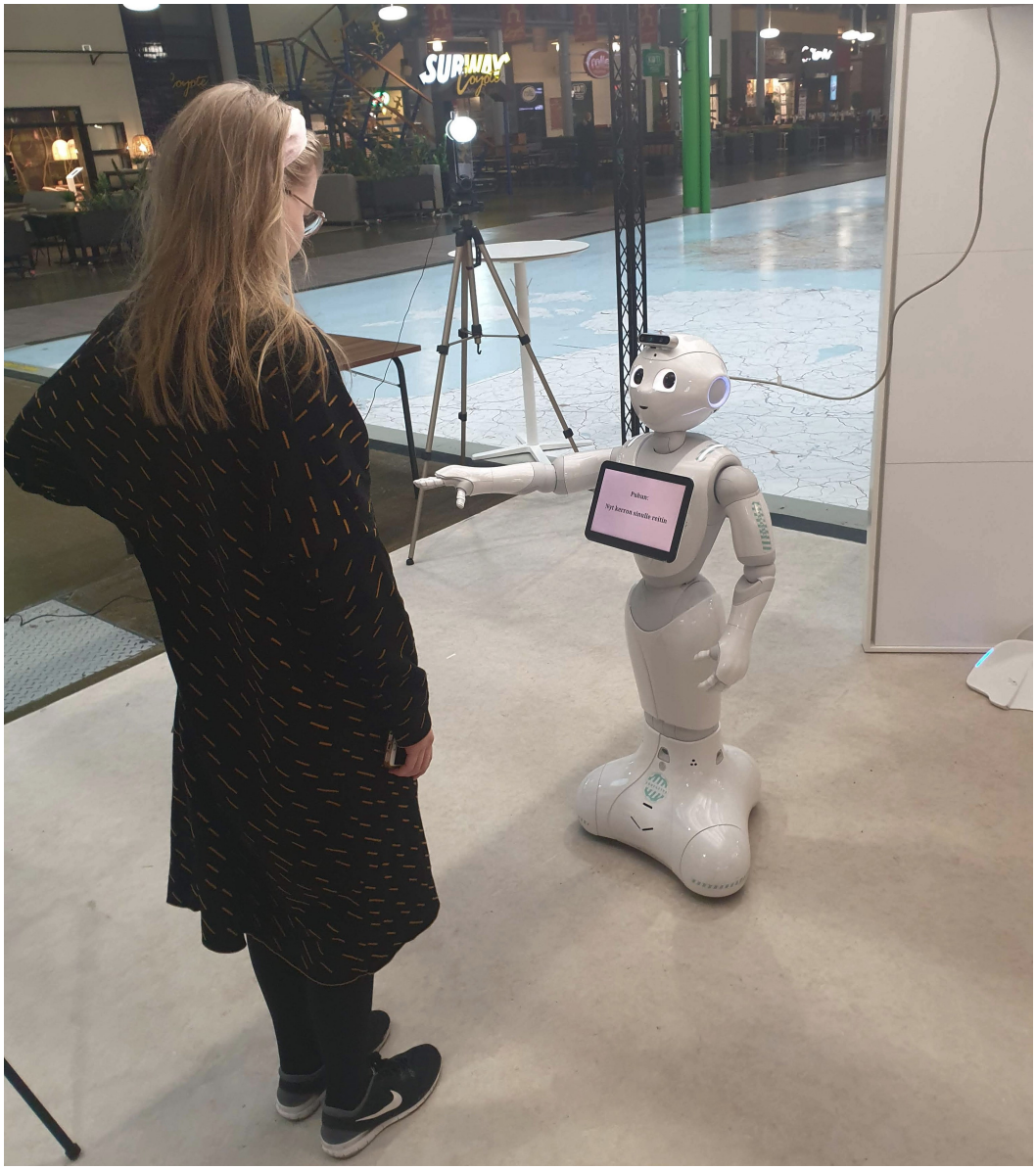


Figure 3.15: A person receiving directions from Pepper. (Image from VTT team)

when it could have moved so they would not risk to be hurt, and to answer people who wanted to know more about the robot or the project than what was explained on the explanatory posters. She was also responsible for starting and shutting down the robot at the beginning and at the end of the half-day. Besides, for security and legal responsibility reasons, we chose to not have the robot navigating during this deployment as it would have been a complicated issue if the robot bumped into someone, especially a kid who could be hurt. It would have been possible if Pepper had a remote emergency stop which could have been given to the guard. Therefore, the step *Ensuring correct placement* was removed in this context. Then, the Human-Aware Navigation component (see Sect. 3.5.6) was disabled and the Shared Visual Perspective Planner (see Sect. 3.5.5) was only used to compute the 360 degrees visibility of a landmark from the person position.

In total, the robot ran the direction-giving task during approximately 96 hours "in the wild". Out of these 96 hours, it was interacting with someone during 45 hours. Table 3.2 summarizes statistical data about the interaction sessions and Table 3.3 summarizes statistical data about the direction-giving tasks.

Description	Value
Number of occurred interaction sessions between a human and the robot	979
Cumulative duration of the interaction sessions	2720 min
Minimal duration of an interaction session	0.1 min
Maximal duration of an interaction session	41 min
Average duration of an interaction session	2.8 min
Standard deviation of sessions duration	3.3 min
Average number of direction-giving tasks during a session	1.1
Percentage of sessions terminated by goodbyes	30%
Percentage of sessions terminated by the participant not perceived by the robot anymore	70%

Table 3.2: Statistics on interaction sessions in the wild

3.7 Integration and test of the QoI Evaluator

As a proof-of-concept for the QoI Evaluator presented in Section 2.5 of the Chapter 2, we integrated it in the direction-giving task described in this chapter.

More specifically, this implementation of the Quality of Interaction Evaluator measured the interaction quality at the direction-giving task level and at the elementary actions level, omitting the interaction session level as this latter was not our focus in the MuMMER project. The QoI Evaluator was integrated into JAHRVIS presented in Chapter 2. The QoI Evaluator is implemented into a Jason function (the reasoning cycle) which is invoked periodically. After multiple testings, we reached the conclusion that it was pertinent, at least in the context of the direction-

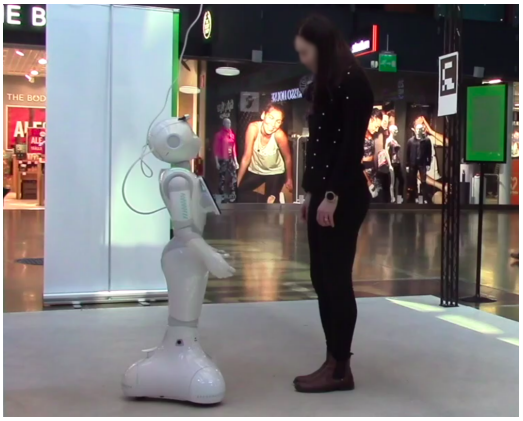
Description	Value
Number of occurred direction-giving tasks between a human and the robot	1156
Cumulative duration of the direction-giving tasks	930 min
Minimal duration of a direction-giving task	0.01 min
Maximal duration of a direction-giving task	22 min
Average duration of a direction-giving task	0.8 min
Standard deviation of direction-giving tasks duration	1.27 min
Success rate of the step *Establishing the shared goal*	63%
Success rate of the step *Route planning according to the human willingness and ability to climb stairs*	100%
Success rate of the step *Pointing to target*	56%
Success rate of the step *Ensuring target seen*	39%
Success rate of the step *Pointing to passage and giving route directions*	94 %
Success rate of the step *Ensuring passage seen or route understood*	92%
Success rate of the removed step *Check if indications understood*	19%

Table 3.3: Statistics on the direction-giving task in the wild. *Ensuring target seen* is a part of the step *Pointing to target* as described in Sect. 3.4.2. Likewise, *Ensuring passage seen or route understood* is a part of the step *Pointing to passage and giving route directions*. The success rate of a step is the number of times the given step has been achieved over the number of times it was planned (e.g. *Route directions and pointing* is not planned if there is no passage to point), all direction-giving tasks combined. Steps were not achieved sometimes because of robot failures but most of the time it was because the human was leaving during the task. As mentioned in Section 3.5.7.1, we did not keep the step *Check if indications understood* all along the deployment because, as shown by the success rate, people were leaving before answering this question. Then, as this step was considered as superfluous by users, we merged it with the one before, *Ensuring passage seen*.

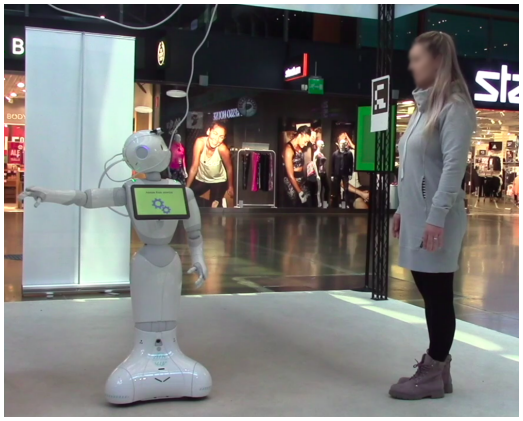
giving task, to have the Evaluator computing the QoI every second for both levels. Therefore, every second, the system computes the value of each metric and then outputs a value for QoI_{task} and QoI_{action}.

(a) A customer listening to Pepper after repositioning

(b) A customer listening and Pepper pointing to a corridor



(c) A customer answering to Pepper



(d) A customer listening and Pepper pointing to a shop

Figure 3.16: MuMMER robot engaged in direction-giving tasks. Around 350 trials with customers in the mall allowed us to gather empirical data to select the metrics and tune the measuring functions parameters.

As mentioned in the step 4b of the chronicle, the robot interacted in the wild with dozens of usual customers (Fig. 3.16), executing around 350 direction-giving tasks. This allowed us to improve the performance of the direction-giving task, to gather standard durations of the subtasks executions and to draw lessons about metric definitions and choices (e.g. we realized it was not relevant to measure the human visual attention towards the robot when it was giving the route explanation as humans look around at this moment). Unfortunately, the practical conditions of the project deployments did not offer us the possibility to evaluate the QoI Evaluator based on a study in the mall with real customers. So, we demonstrated

– after improvements of the metrics equations such as the Distance-to-Goal one, and manual tuning of their parameters based on the experience in the mall – our finalized concept through tests in our lab (step 6). This is shown in Sect. 3.7.3 where we present and discuss, a comparison of the QoI computed by the robot when it is dealing with an "ideal" human, a "confused" human and a "non-compliant" human during a direction-giving task, performed in the lab. Before that, we present in Sect. 3.7.1 and Sect. 3.7.2 how the QoI is evaluated at both task and action levels for the direction-giving task.

3.7.1 QoI Evaluation at the task level

In the context of the direction-giving task, we have selected two metrics to evaluate the QoI at the task level: a metric defined in the Sect. 2.5.4, the *Deviation from standard duration* and, the aggregation over time of the actions QoIs. Following the process of Fig. 2.2, we measure the QoI of the Task$_i$ = direction-giving_task, based on the QoI of all task actions and Task metric$_1$ = Deviation from standard duration.

The *Deviation from standard duration* is used to measure the QoI at the task level as the task is a sequence of subtasks. Indeed, if the subtask lasts longer than expected, the QoI should decrease. Then, as needed for the metric computation we have determined the values of the soft deadlines SD_i for each subtask $a_i, i \in [0, 4]$, using the empirical data we gathered as explained in. Specifically, we have computed the average time execution of each subtask, after removing the cases for which the execution of the subtask was annotated as not smooth. These soft deadlines are presented in table 3.4. Finally, we chose $V_i = 0.5$ for all the subtasks.

ref sec

Subtasks	soft deadline (s)
Target refinement process	30
Ensuring Correct HR Placement	30
Ensuring target seen	20
Direction explanation and pointing	30
Ensuring Direction Seen	20

Table 3.4: Soft deadlines SD_i for each subtask of the direction-giving task

The task QoI is also dependent on the actions QoI values (their computation is described in Sect. 3.7.2). Indeed, the actions QoIs should be reflected on the task QoI as, if a majority of the actions have a low QoI, the task QoI cannot remain high. That is why, besides the *Deviation from standard duration*, we take into account the average of the action QoIs of the actions already executed or still running.

Then, the task QoI is computed using Equation (2.1) presented in Sect. 2.5.3. After various trials we have empirically chosen the weights W_i for each metric $M_i, i \in [0, 1]$. The final equation to compute the task QoI is:

$$QoI_{dir-giv\_task}(t) = \frac{\Phi_{dir-giv\_task}(t) + 3 * \overline{QoI}_{actions}}{4}$$

3.7.2 QoI Evaluation at the action level

As mentioned earlier, each subtask of the direction-giving task can be decomposed into actions. These actions involve several turn-taking steps, the robot asking complementary information, informing the human or expecting an action or reaction from them. We need to measure the QoI during the execution of each action. To do so, we have chosen one or more metrics for each action.

Metric id	Metric name	Metric equation – with Equations of Section 2.5.4	Scaled metric – with functions
$M_{H\_contrib}$	Human contribution to the goal	nb_R_repet	$n_1(nb\_R\_repet) = 2 * \frac{nb\_R}{}$
M_{Exp_SI}	Fulfilling robot expectations about social interaction	$Ar = \frac{duration_{isAttentiveTo(robot)=true}}{duration_{robot\_speaks}}$	$n_1(Ar) = 2 * Ar -$
M_{DtG}	Distance-to-Goal	$\begin{cases} \Delta DtG(t=0) = 0 \\ \Delta DtG(t) = \max(0, \Delta DtG(t-1) - 1) \\ \quad \text{if } path\_length(t) < path\_length(t-1) \\ \Delta DtG(t) = \Delta DtG(t-1) + 1, \text{otherwise.} \end{cases}$	$-s_1(DtG(t)) = -1 + 2\exp\left(-\ln\right.$
M_{TtG}	Time-To-Goal	$\Delta TtG(t) = \max(0, e(t) + TtG(t) - TtG(T_0))$	$-s_1(TtG(t)) = -1 + 2\exp\left(-\ln\right.$

Table 3.5: Metrics used in the implementation presented in Section 3.7.

For each action of the following list, we explain which metrics M of Table 3.5 we have used and scaling functions of Appendix A and then, how we compute the action QoI.

(a) *Robot-Human information sharing:* The robot speaks to the human, shares information such as the route direction and announces the next steps of the plan. The robot expects that they are paying attention to it. Therefore, we use the *Fulfilling robot expectations about social interaction* M_{Exp_SI} based on the attention ratio. Two parameters need to be defined for the scaling function, the bounds b_1 and b_2. As the minimum value for the metric, a ratio, is 0 and the maximum value is 1, then $b_1 = 0$ and $b_2 = 1$. The QoI of the action is computed with this only metric.

(b) *Human-Robot Q/A process:* The robot asks a question to the human. As for the previous action, the robot expects the human to pay attention to it

so we compute the QoI with M_{Exp_SI}. It also expects the human to give an appropriate answer. If it does not happen, it will ask the human to repeat, specifying that the answer has not been understood. We have limited the possible number of attempts to 3. After 3 attempts, the robot ends the task, as it cannot carry on with the task without an answer. So, we use *Human contribution to the goal* $M_{H\_contrib}$, the number of times the robot repeats. Because the maximal number of repetitions is 3, we set for the scaling function $b_1 = 3$ and $b_2 = 0$.

The QoI is computed with the two metrics: *Fulfilling robot expectations about social interaction* and *Human contribution to the goal*. The trials showed that the action QoI results were satisfying with the weights $W_i = 1, i \in [0, 1]$ as applying the Equation (2.1).

(c) *Ensuring that Human moves aside:* This action is used if, for pointing, the robot decides to place itself in a position which is very close to where the human is currently standing. In this case, the robot asks the human to step aside to the right or left, depending on the human's future position. Then, we want to measure the progress of the human going further from the planned robot position. In order to do this, we use the *Distance-to-Goal* M_{DtG} but with the condition of the ΔDtG equation adapted, being if $path\_length(t) > path\_length(t-1)$ instead of if $path\_length(t) < path\_length(t-1)$. We scale the metric with $-s_1$, the additive inverse of the scaling function and not directly s_1 as the closer to 0 ΔDtG is, the better it is in terms of goal completion. From trials, we set $-s_1$ parameters values with $th = 5$ and $k = 1.5$.

If the human does not move or does not go far enough from the robot position, the robot will ask again with a limit of 3 trials (if the robot cannot move, it will carry on the task from their current positions). So, we use $M_{H\_contrib}$ as for the previous action.

(d) *Human-aware robot navigation:* The robot has to move from its initial position to its computed one. It navigates while respecting social constraints and its path may change as it adapts according to what the human is doing. At execution time, to measure the robot progress towards its goal, we use the *Time-to-goal* M_{TtG}, with the same scaling function than M_{DtG}. The QoI of the action is computed with this only metric.

(e) *Ensuring correct human placement for verbal interaction:* After it has moved, the robot asks the human to come in front of it. If the human is not perceived after a few seconds, the robot will ask again and so on in a maximum of 3 trials. If after these 3 times the human is still not perceived, the robot ends the task.

The QoI of this action is computed with $M_{H\_contrib}$ – we do not use M_{Exp_SI} as the human is not in the field of view when the robot is calling them.

(f) *Ensuring correct human placement for route explanation:* Once the human is in the robot field of view after the HR motion, they may not be at the right place to properly see what the robot has to point at. In this case, the robot will ask the human to move forward or backward according to what it has computed about the human perspective (e.g. this is to avoid that an object occludes the view for the human). Then, we want to measure the human progress towards the position the robot has computed for them. In order to do this, we use the *Distance-to-Goal* M_{DtG}.

The robot stops giving instructions if it computes that the position of the human allows them to see the target, or after 3 trials, so we use $M_{H\_contrib}$. After 3 trials, if the human cannot see the target, still, the robot will carry on the task taking this into account.

Mall elements	Mockup mall	Real mall
Shops	19	140
Doors, stairs, elevators	10	50
Corridors	11	41
Levels	2	2

Table 3.6: Number of elements described in the mockup and real malls (geometric, topologic and semantic models in Fig. 3.5).

3.7.3 Proof-of-Concept

This section reports on an effective implementation of the approach as an illustrative proof of concept. We show the ability of the robot to conduct an interactive task, to assess in real-time the QoI and to track its evolution during three direction-giving task executions where the same human displayed a different way of behaving. In the three cases, the task was conducted until its end, in our lab where we reproduces the mall environment (Fig. 3.5a, Table 3.6). The computed QoI for each way is presented in Fig. 3.17. The three different ways of behaving are described in the following list:

- A human executed perfectly the expected actions and was not disturbing the robot when it navigated (i.e. the "ideal" human from the robot point of view).
- A bit "confused" human tried to contribute to the task success but did not execute everything well. The human was, from time to time, not very attentive, as looking around. Also, they gave an answer to the first question that the robot did not understand, and then they took their time before answering again. Then, they prevented a bit the robot to move as it had planned and once the robot reached its position, they took time to come as close as the robot wanted.

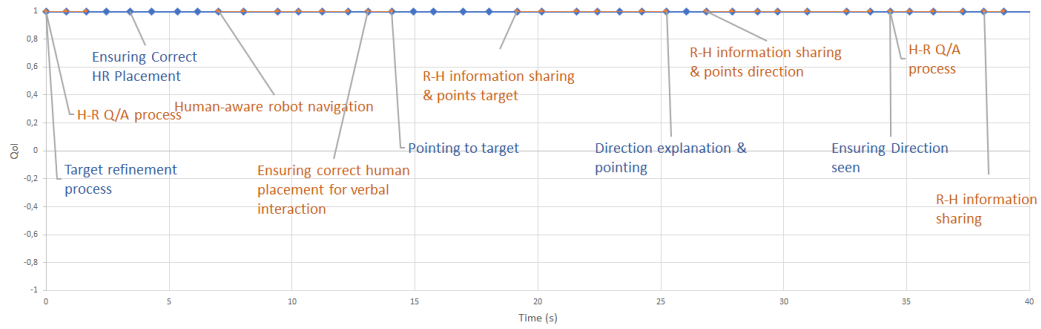
Action	QoI formula (metric aggregation)
Robot-Human information sharing	$M_{Exp\_SI}(t)$
Human-Robot Q/A process	$\frac{M_{Exp\_SI}(t) + M_{H\_contrib}(t)}{2}$
Ensuring that Human moves aside	$\frac{M_{DtG}(t) + M_{H\_contrib}(t)}{2}$
Human-aware robot navigation	$M_{TtG}(t)$
Ensuring correct human placement for verbal interaction	$M_{H\_contrib}(t)$
Ensuring correct human placement for route explanation	$\frac{M_{DtG}(t) + M_{H\_contrib}(t)}{2}$

Table 3.7: QoI computation for each action as an aggregation of metrics

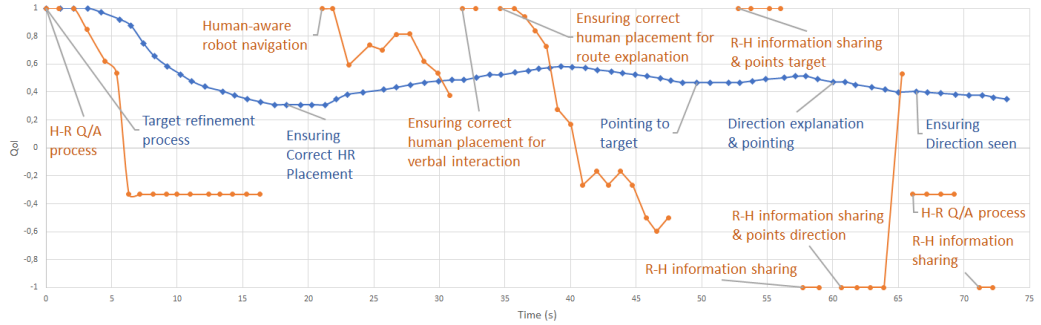
- A human wanted to disturb the robot during the task. They gave three incomprehensible answers to the first question, blocked multiple times the robot in its move, waited for the robot to ask twice to come in front of it and finally asked the robot to point and explain the route three times.

Now, if we take a look at the QoI outputs of Fig. 3.17, we can see that their three shapes are very different. In Fig. 3.17a, we can observe that the task and actions QoIs remain with the highest value 1 all along. A graph as this one allows us to infer that everything went very smoothly during this direction-giving task. Then, we can guess that it corresponds to the execution performed with the 'ideal' human.

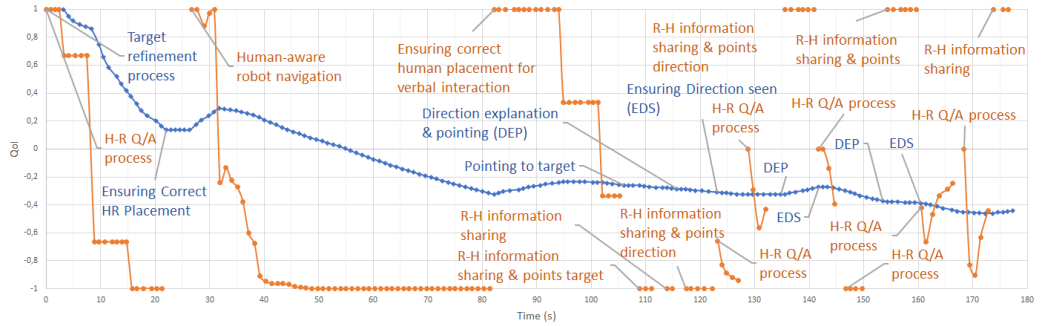
In Fig. 3.17b, we note that each subtask was executed in respect of the standard duration. If the QoI of *Target refinement process* drops it is because of the action QoI as the QoI of the *H-R Q/A process* drops because the robot had to repeat the question and the human was not looking at it. From 21 seconds to 40 seconds, we can see the task QoI getting higher as the QoIs of *Human-aware robot navigation*, *Ensuring correct human placement for verbal interaction* and the beginning of *En-*



(a) Evolution over time of the measured QoI for the 'ideal' human. Both action and task QoIs remain at 1 as the task is proceeding smoothly.

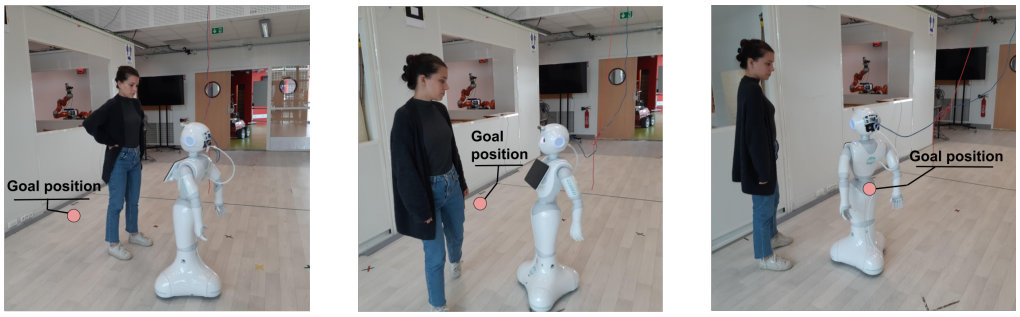


(b) Evolution over time of the measured QoI for the "confused" human. They took time to answer the first robot question and to move forward but the task QoI does not drop too much because the robot was able to give the route explanation without any issue even though the human was not very attentive.



(c) Evolution over time of the measured QoI for the non-compliant human. Several times the human did not give the expected answer to the robot during the target refinement process. Then, they blocked the robot path. After that, the robot had to ask twice the human to come in front of it. Finally, the robot repeated the route direction three times but still the human kept saying that they did not understand. Therefore, the task QoI decreases all along the task.

Figure 3.17: Evolution over time of the measured QoI for the route guidance task with three different human behaviors. The QoI for the task is drawn in blue, and the QoI for the actions is drawn in orange.



(a) Human who put themselves on the robot path, preventing the robot to navigate towards its goal position

(b) Human who put themselves on the robot path after it computed a new path to reach its goal position

(c) Human finally getting outside of the robot path, allowing it to reach its goal position

Figure 3.18: A human disturbing the robot during *Human-aware navigation*, preventing it to reach its goal position as planned.

suring correct human placement for route explanation are quite high. Next, seeing the shape of the computed QoI of the action *Ensuring human placement for route explanation*, we can infer that the human was not moving as the robot wanted. Indeed, they took 10 seconds to make one step forward (they had 1 meter to cross). Because of that, the task QoI started to decrease again. In the final part of the task, the human was time to time attentive to the robot answered quickly to the last question, so the task QoI remained rather equal with its final value being 0.34 which is above 0 so meaning a correct interaction.

Finally, we can see in Fig. 3.17c that the final QoI of the task is -0.44 which allows us to infer that the task was not executed smoothly. And indeed, when we look at the shape of the task QoI, it only went down (or almost) all along the task. It is explained by some subtasks that took more time than they should have and also by some actions QoIs that are very low, especially the one of *Human-aware robot navigation*. At the beginning of the robot navigation, the estimated time to goal returned by the planner was 6 seconds but the robot actually took 50 seconds to reach its goal then the action QoI computed with $= M_{TtG}(t)$ was -1 for 40 seconds. And indeed, all along its navigation, the human was blocking the robot until they got tired of this game, as visible on Fig. 3.18.

In this example, we showed the QoI evaluation process integrated to a complete robotic architecture. The robot was able to assess the QoI in real-time while interacting with a human.

3.7.4 Discussion on the results of the QoI Evaluator

While a number of evaluation methods has been proposed to evaluate a human-robot interaction from the human perspective and often for analysis after performance, our choice to let the robot evaluate, on its own and in real-time the quality of its interaction with a human is quite new and original. To endow the robot with such

an ability, we designed, implemented and tested a number of metrics and a method to aggregate them.

The work of Steinfeld *et al.* [Steinfeld 2006] was very helpful to design a first set of metrics and as an inspiration about what could be used. From there, we have elaborated and proposed a set of metrics which are meant to estimate of the quality of an ongoing interaction and not once it is over. The work of Hoffman [Hoffman 2019] regarding the *fluency* definition and how to measure it was also inspiring. In a way, we extended his work by giving a meaning to the fluency measurement on the robot side, and in real-time – while their work applies to offline evaluation of shared workspace tasks. In Sect. 2.5.2, we mentioned systems measuring human affective states in real-time such as the framework developed by Tanevaska *et al* [Tanevska 2017]. Although we think such metric could be an interesting additional information to assess if an interaction is going well, we believe that these measurements do not offer an accuracy that would lead to objective measurement of the quality of interaction, thus, we did not introduce them in our set for now. However, this could be done since our framework is designed to be open to new metrics. As for contributions, like the one proposed by Anzalone *et al.* [Anzalone 2015], based on metrics such as gaze, head pose, body pose and response times to measure real-time engagement, we took them into account to some extent. However, the measure of the engagement that we propose should be refined depending on the inputs available on-line to the robot . Moreover, we will investigate how their work could be used in a more general way (e.g. depending on the action that should be done and its context, human head pose and body posture could be a good indicator of effectiveness and not only engagement).

Our intention, when we developed the idea of the Quality of Interaction Evaluation, was to use such computation to feed the decision-making process of the robot and this is what we intend to do in the future. However, such framework can also be used to compare interactions between different humans and/or robots, eventually as a benchmark similarly to the work of Sanchez-Matilla [Sanchez-Matilla 2020] or as a way for developers to detect repetitive interaction issues with an unsupervised robot in a real-world environment.

As a proof-of-concept, we implemented and deployed a first version of a QoI Evaluator assessing task and actions QoIs. We tested it on an interactive robot dedicated to provide route guidance to customers in a large mall. The approach gave satisfactory results. It showed the potential ability of a robot to detect momentary decreases of the Quality of Interaction and also more serious degradation of it which may need drastic change of behavior for the robot. This is only a first step and it should be validated with a study where we will ask humans to evaluate the quality of their interaction with the robot in a similar manner. The goal will be to analyse and compare this to the evaluation of the interaction quality estimated by our robot and, based on that, investigate potential improvements.

Finally, we do not claim to have a perfect measure of the Quality of Interaction. However, although the concept of Quality of Interaction is quite abstract, Movellan *et al.* showed that when it is measured by human observers, the inter-observer

reliability of the concept is quite high. Therefore, we believe we can endow the robot with an effective and pertinent ability aiming at measuring the quality of an interaction. We are aware that the set of metrics we proposed to do so is not exhaustive but the framework is designed to be easily extended with new metrics.

3.8 User Study

Chapter 4

The Director Task: a Psychology-Inspired Task to Assess Cognitive and Interactive Robot Architectures

Contents

In this chapter, we propose a new psychology-inspired task, gathering perspective-taking, planning, knowledge representation with theory of mind, manipulation, and communication. Along with a precise description of the task allowing its replication, we present a cognitive robot architecture able to perform it in its nominal cases. In addition, we suggest some challenges and evaluations for the Human-Robot Interaction research community, all derived from this easy-to-replicate task.

The contribution presented in this chapter is excerpted from our work, published in the proceedings of the RO-MAN 2021 conference [Sarthou 2021]. This contribution closes this thesis and has been achieved in collaboration with other PhD students of the HRI teams. Guilhem Buisan was concerned about the task planning part. Amandine Mayima worked on the supervision component. Kathleen Belhassein has designed the presented task with us giving her psychologist point of view to create a task on which user studies could be performed. The engineer Yannick Riou worked on the motion planning component allowing us to develop a task where the robot acts on its environment. My concern about this task has been the integration of my previous contributions about Ontology and the REG. It has also been the opportunity to create an entire architecture extending the ones presented all along with this thesis and linked with the contributions of the team. Finally, I contribute to the Situation assessment component and on the Language understanding part.

The components related to my teammates will be briefly described to give an overview of the architecture. The newly introduced capabilities on which I work will be more detailed to explain the links I make between all my contributions, centred on the knowledge representation.

4.1 Introduction

Developing robotic architectures adapted to Human-Robot Interaction and thus able to carry out interactions in an acceptable way is still today a real challenge. The complexity comes, among other things, from the number of capabilities that the robot must be endowed with and therefore from the number of software components which must be integrated in a consistent manner. Such architectures should provide the robot with the capability to perceive its environment and its partners, to merge and interpret this perceptual information, to communicate about it, to plan tasks with its partner, to estimate the others' perspective and mental state, etc. Once developed, the evaluation of these architectures can be difficult because all these components grouped into a single system. The tasks we usually want the robot to handle must highlight a maximum of abilities, while still being simple enough to be reproduced by the community. Moreover, we should be able to conduct user studies with it to validate choices regarding naive users.

Since a long term goal of the robotic field is to see robots evolving in our daily life, many tasks and scenarios have been inspired by everyday activities. Even if these tasks offer a large variety of situation to be handle, since the human partner is not limited in his actions, they have the disadvantage of not highlighting some subtle abilities which are nevertheless necessary for good interaction. The robot guide task [Satake 2015b] in mall, museum, or airport, requires high communication skills to understand free queries (possibly involving chatting) and respond to them, whether to indicate a direction or to give advice. However, the perception needs can be limited due to the vast environments, as well as the perspective-taking needs due to

the same perception of the environment by the robot and the human[1]. Finally, with such a task the human partner is not an actor of the task and just has to listen to the robot once their question is asked. Even if being in more constrained environments, bartender-like tasks [Petrick 2012] have the same disadvantages. Indeed, the human is considered as a customer, and as such, the interaction with the robot is limited. The robot will never ask the human to help it for performing a task and their actions do not require coordination either full collaboration.

To involve the human partner in the task and requiring him to act with the robot, assembly-like tasks [Tellex 2014] can be used. Nevertheless, in most cases, the human acts as an assistant rather than as a partner as full collaboration can be challenging to perform. The robot thus elaborates a plan and performs the assemble, then asks for help when detecting errors during the execution (e.g., when it cannot reach some pieces). Here the task leads to unidirectional communication. Moreover, because in such a task both the robot and the human have equivalent knowledge about the environment, it can be hard to design situations where belief divergence appears and thus perspective-taking would be required.

Scaling down an everyday task to transform it into a toy task around a table can reduce the task complexity and allow easy reproducibility. Moreover, it allows the robot and the human to work in the vicinity of each other, with smaller robots for example. With the toy version of the assembly task presented in [Brawer 2018], the human is more involved in the task. They ask the robot to take pieces and to hold them to help them assemble a chair. Even if the communication is unidirectional, we could imagine inverting the roles to test different abilities. Moreover, communication implies objects referring with the use of various visual features about the entities. Even if both agents have the same knowledge about the environment, the communication is grounded according to the current state of the world. In this task, no decision has to be made by the robot but once again, inverting the roles could open other challenges.

To focus studies around perspective-taking and belief management, the Sally and Anne scenario, coming from a psychology test, has been studied in robotic [Milliez 2014b]. In this scenario, the robot is an observer of a situation where two humans come and go from a room, and move an object from a box to another. Since a human is in the room when the other is acting, a belief divergence appears between the two humans and the robot has to understand it. While the task highlights the belief management, it is first limited regarding the perspective-taking since the human presence or not could be sufficient to estimate the humans beliefs[2]. Moreover, the humans do not act with the robot since it is just an observer of the scene. In addition, no goal is formulated and the human neither interacts with one another. Finally, no communication is needed in the task. The scenario

[1] For sure we can find some tricky cases where it could help but they do not reflect common situations.

[2] When both humans are in the room they have the same perception of the scene but have different beliefs about hidden objects. Perspective-taking would be required if the humans could lean over the boxes to check what is inside.

is thus focussed on the analysis of a situation.

In this chapter, we first propose a new psychology-inspired task that we think to be challenging for the Human-Robot Interaction community and rich enough to be extended: the Director Task. Inter alia, it requires perspective-taking, planning, knowledge representation with theory of mind, manipulation, communication, and decision-making. Then, we present the robotic cognitive architecture that we develop to perform the task in its nominal cases. Finally, on the basis of the presented task and what has been developed, we present a discussion about the possible future challenges and evaluations for the research community, with possible extensions of the task.

4.2 The Director Task: From psychology to Human-Robot Interaction

In this section, we present the origins of the Director Task and the needs it aims to respond to regarding other tasks from the psychology. We then detail the setup we have designed in terms of objects characteristics and organisation in the environment. We end this section with our adaptation and the required abilities we have identified.

4.2.1 The original task

The Director Task has been mainly used in psychology as a test of the Theory-of-Mind usage in referential communication. This task originates from a referential communication game from [Krauss 1977]. In this game, two participants are one in front of the other with an opaque panel between them. A speaker has to describe odd designs to a listener, either to number them for the adults or create a stack of cubes for the children. To refer to the odd figures, participants have to use images (e.g. “it looks like a plane”).

This game was then adapted by Keysar et al. [Keysar 2000] and becomes the Director Task. It has been used to study the influence of mutual knowledge in language comprehension. In this task, two people are placed one in front of the other but instead of an opaque panel between them, they place a vertical grid composed of different cells and objects in some of them. The **director**, a participant or in most cases an accomplice, instructs the **receiver**, a participant, about objects to move in the grid. The receiver thus follows the director’s instructions about objects to move. The particularity of the task is that some cells are hidden from the director, meaning that the receiver, being on the other side of this grid, does not have the same perspective as the director. He thus knows the content of more cell than the director and consequently sees more objects. When the director instructs the receiver to move an object, for a successful performance, participants must take the perspective of the director to move the right one. Because the configuration evolves all along with the task, he has to update this estimated perspective all along with

the interaction.

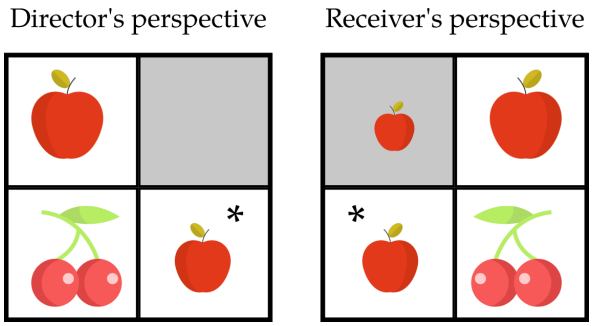


Figure 4.1: Sample display from the director's and the receiver's perspectives. The asterisk indicates the target object. Giving the sentence "the smallest apple" the receiver should find the good one even if he can see a smallest one in its perspective.

Taking the example of figure 4.1, if the director instructs the receiver to take the smallest apple, the target object in its perspective is the one marker with the symbol *. However, for receiver, in its perspective, the target object is not the smallest apple since the smallest one (called distractor) is only visible by the participant and not by the director. The participant then must understand the director's perspective to take the target apple and not the distractor. Some studies showed that for their first attempt, participants took the smallest apple from their own point of view and only after, the target one. These results were interpreted in [Keysar 1994, Keysar 1998, Keysar 2002, Keysar 2003] as the participants understanding language in an egocentric way. Some social cognition studies used a computer-version of the Director Task [Dumontheil 2010] whose results are consistent with the ones mentioned previously, namely that participants do not use Theory-of-Mind inferences in language interpretation.

Although Theory-of-Mind and perspective-taking both require the attribution of mental states to others, some authors trend at distinguising Theory-of-Mind tasks and perspective-taking tasks as involving distinct although related mechanisms. In [Santiesteban 2012], they consider that perspective-taking abilities were measured by the Director Task whereas Theory-of-Mind usage was investigated through another task called "strange stories" [Happé 1994]. This Theory-of-Mind task requires the attribution of mental states to a story protagonist, meaning to maintain an estimation of others' mental states. At the difference, the Director Task requires the adoption the perspective of the director in order to follow the instructions, meaning to use this knowledge in order to execute the task properly. In this way, the authors estimated that the Director Task requires a higher degree of self-other distinction by continuously isolating our own perspective from the director one, in order to use it to act. In addition to perspective-taking abilities, the Director Task makes use of executive functions [Rubio-Fernández 2017] (i.e. vary the processing of information according to current goals in an adaptive manner) and attentional resources [Lin 2010].

To summaryse, the Director Task has been used to study referential communication, language comprehension, and perspective-taking abilities. However, to our knowledge, it has never been exploited in the context of a HRI although this task presents interesting challenges for this field. More than technical challenges, it provides a way to investigate the different cognitive and behavioral processes involved in such a cooperative Human-Robot task.

4.2.2 The Director Task setup

The material used in this task has been chosen to be easily acquired and can be hand-built. It is composed of blocks, compartments, and a storage area. Each element is equipped with AR-tags allowing the robot to perceive them without advanced perception algorithms.

Figure 4.2: Part of the material used for the Director Task. Each element is equipped with AR-tags allowing their detection by the robot. Each block has four visual characteristics: a main color, a border color, a geometric figure and a figure color.

Three types of compartment exist and are illustrated on the right part of figure 4.2. The basic ones are open on two of their opposite sides (d). They allow both the receiver and director to see the content and to manipulate it. Others are open only on one of their sides (e). With such a compartment, only one of the participants can see and take what is inside. The other participant can neither know if a block is inside or not. The last compartment type (not used in the implemented version) has an open side and the opposite one equipped with a wired mesh (c). Because of the side with the wire mesh, both participants can see what is inside but only one of them can take it. Thanks to these three types, we will be able to vary the awareness of the blocks (e.g., a block is known to be present but not necessarily visible), the visibility of the blocks, and their reachability (e.g., a block can be visible but not reachable). While the original Director Task uses a vertical grid, we prefer here to use several compartments to create the grid. It allows more modularity to create different situations.

While the tasks used in psychology use everyday objects, we rather choose blocks that can easily be manipulated by robots and on which we can fix tags for their detection (a-b on 4.2). The blocks have a primary color covering them all. On two

opposite faces, additional visual features are drawn. The top part of these faces is dedicated to the robot's perception with a unique AR-tag on each face[3]. The bottom part is the same on both faces and is dedicated to human perception. In addition to the primary color, three visual features are available for the human to distinguish them: a colored border, a colored geometric figure (both the color and the figure can change making two features). Every visual feature (the colors and the forms) has exactly two variants. The colors are either blue or green and the figures are either a triangle or a circle. We can thus have 16 unique blocks.

The agents can use the four visual features to refer to a specific block and the complexity of the description depends on the used features. While the main color is directly related to a block, the other colors are respectively related to the border and the figure. In this way, for two blocks for which the only difference is the color of one of these elements, the said element has to be referred to in order to refer to the divergent color. A description of a block involving all its four features would be "the [color] block with the [color] border and the [color] [figure]".

The figures and colors have been chosen in such a way to allow the emergence of "coded words" between the participant to identify a block. With a bit of imagination, some could refer to the left-most block (a) through the sentence "the mountain in the sea" or the other (b) by "the puddle".

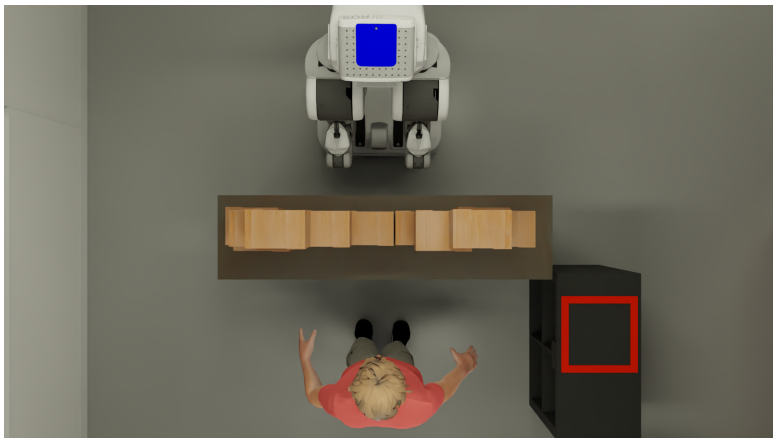


Figure 4.3: The Director Task setup with the robot and the human partner one in front of the other and a piece of furniture between them. Compartments are placed on top of the furniture and blocks are placed in the compartments. Next to the agent having the receiver role, here the human, a storage area is placed to drop the removed blocks.

Regarding the disposition, the compartments are stacked on a piece of furniture to create a kind of grid. The blocks can be put in a compartment. As illustrated in figure 4.3, the two agents are placed one in front of the other with the furniture and thus the compartments between them. Finally, one storage area, corresponding to

[3]Since the tags are different on each side, the director cannot refer to them as the receiver does not see the same ones

the place where the receiver has to store the blocks, is delimited by a rectangle on a shelf next to the receiver. In the figure, the human would be the receiver since he has the storage area on his right.

4.2.3 The adapted task

Now we explained the Director Task setup and the available material, we present the rules we have adapted for HRI applications. First, the high-level goal of the task is known by both agents: to put a set of blocks away. The precise goal is given by the experimenter to the director, either the robot or the human. It corresponds to a subset of the blocks presents in the compartment that the receiver should remove from and put in the storage area. This choice, to remove the objects instead of moving them in the grid, enables an evolution of the situation over time. It thus requires a constant adaptation during the interaction. The goal can be given on a sheet of paper, a screen behind the receiver, or marks on the blocks on the director side. No block order is required in the formulation of the goal. The director is thus allowed to elaborate a strategy if needed.

As mentioned previously, the Director Task characteristics bring a number of interesting challenges for a collaborative robot to solve. Because this is a task with roles, one of the first challenges is to build a robotic architecture that gives the robot the ability to play both roles. Then, each role brings some specific problems to solve from a robotic point of view.

In order to enrich the task with perspective-taking, we adapted the task so that both the director and the receiver have to use perspective-taking. Since in the original task, the director knows he has a subset of the receiver's perspective, he can consider all the objects when communicating. Thus, only the receiver has to reason about the other's perspective, taking into account that some objects are not visible by the director. For HRI applications, we use the one side hidden compartments in a way to also have objects hidden from the receiver and visible by the director. Therefore, both roles have to perform perspective-taking, whether to give instructions or to understand them. On the illustration of figure 4.4, the director (left image) can instruct the receiver to take the blue blocks as the other blue blocks in his perspective is hidden from the receiver. From the receiver point of view (right image), he can find the instructed block as the other blue block is hidden from the director.

To be able to study precise skills, such as verbal communication, perspective-taking, and adaptation, we defined a set of rules for both roles. First, the agents are not allowed to point to objects, either with their hands or gaze. They thus have to verbally describe the objects, focusing the task on verbal communication. However, to avoid too easy description of the kind "the fully green block", we remove the four uni-color variants[4]. In addition, to not fall into simple referential communication task, participants are not allowed to use spatial relations in the

[4]When we said too easy it is from the human point of view, generating and understanding such description can be challenging for a robot.

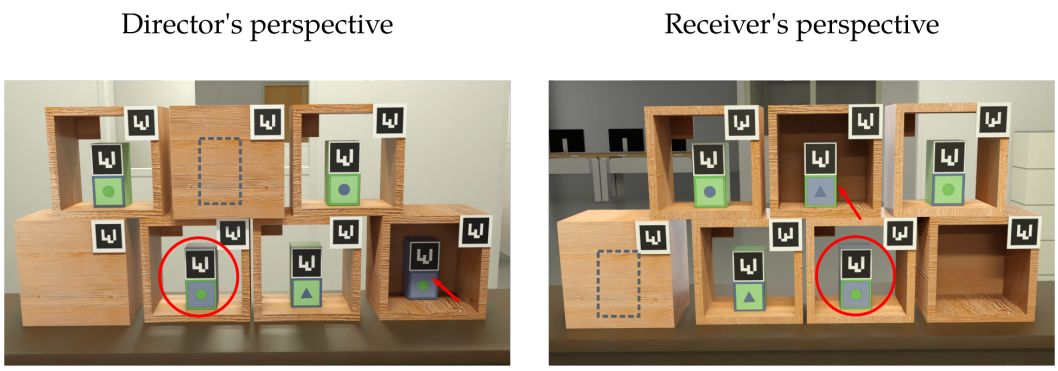


Figure 4.4: A director task setup adapted to the HRI with the director's and receiver's perspectives. For the material, each element (blocks and compartment) is equipped with AR-tags allowing their detection by the robot. Each block has four visual characteristics: a main color, a border color, a geometric figure, and a figure color. Compartments can be hidden for the director or the receiver. For the director to designate the block marked with a red circle, estimating the receiver's perspective, he can refer to it by its main color (blue) because he estimates the other blue block is not visible by the receiver. For the receiver, by taking into account the director's perspective, he can understand the referred block as he estimates the other blue block to not be visible by the director.

verbal communications. They cannot, for example, say "the leftmost block" or "the block to the right of the green one". In this way, they are limited to few visual features, with high ambiguity. Since a description of a block using its four visual features can be hard for the human to process, we first expect the participants to minimize the complexity of their communication by referring to the blocks only using the features distinguishing them from other blocks. Moreover, we also expect the participant to take into account the other perspective allowing once again to minimize the complexity of the communication.

Over these elements, we can see that the task can easily be replicated and offer a controlled setup, making it a good task for human-robot user studies. Moreover, due to the number of involved processes and the number of situations that can be made, there are a lot of elements that can be analyzed and explored. Also, with the same setup, it is possible to perform human-human studies or human-robot studies which can be interesting to compare.

4.2.4 Additional abilities

More than being an easily reproducible scenario to perform user studies on human-robot interactions in a controlled environment, the Director Task allows to demonstrate abilities of a robotic system. We detail here some additional abilities for which the task has been designed for.

Planning When a large number of blocks has to be considered to achieve the goal, it quickly becomes complicated to communicate about some of them as the director would have to add a lot of adjectives to be able to refer to one block. Therefore when the robot is the director, it becomes interesting to integrate the communication and the task planning. Indeed, depending on the order in which the blocks are designated, the complexity of instructions can decrease or increase over time. Then, the planner can return an optimal order in which the robot has to give the instructions to the human.

Contingencies handling While performing the Director Task, errors can easily happen. Either because the director gives a wrong instruction or the receiver misunderstands the instruction and takes the wrong block. In both cases, it can be because of a wrong consideration of the other agent's perspective or simply inattention. Moreover, because some instructions might be right but hard to interpret by the receiver leading also to an error from them. Finally, errors can happen because of failure of the robotic system, as a failed action execution leading to a block to fall on the floor. A robot with a robust decision-making system will be able to analyze, try to determine their origin, and handle a number of these contingencies. For example, if the human takes the wrong block, the robot can react in different ways, either by asking the human to put it back if this block is not part of the goal, or saying nothing and re-planning if this block was among the ones to take. If errors happen repeatedly, the robot can also react differently than for a punctual error and maybe try to modify its behavior.

Communication We saw that the task requires to put a focus on communications. The communication about an object can be more or less efficient, depending on the number of characteristics given about the object or the pertinence of these characteristics. Instructing for the blue block with a circle in figure 4.4, the geometrical figure information is not mandatory. Thus, the robot needs to be able to give proper instructions but also to understand the human ones. Moreover, in complementarity with the error management, the robot can communicate to help to solve the detected contingency. Taking a situation (with the disposition of figure 4.4) where the human as director instructs the robot to remove the green block with a circle. This instruction matching two blocks, the robot could say that it does not find the instructed block. A preferable reaction would be to help the human to refine the instruction and say "the one with a blue circle or a green circle ?".

4.3 The cognitive architecture and the knowledge link

In this section, we present the architecture developed to handle the Director Task in its nominal case for both roles. The architecture aims at being extending but already endows the robot with the abilities listed previously even if there are not mandatory to achieve the task. This architecture is the continuity of the one presented all along

this thesis. It can also be seen as a whole new instantiation of the deliberative architecture for Human-Robot Interaction presented in [Lemaignan 2017a]. The seven identified modules are represented in figure 4.5 with their respective communication links. In the rest of this section, we detail each module and how we have refined them in terms of functionality and links to thers. The modules already presented in this thesis will be briefly recalled but not detailed in depth.

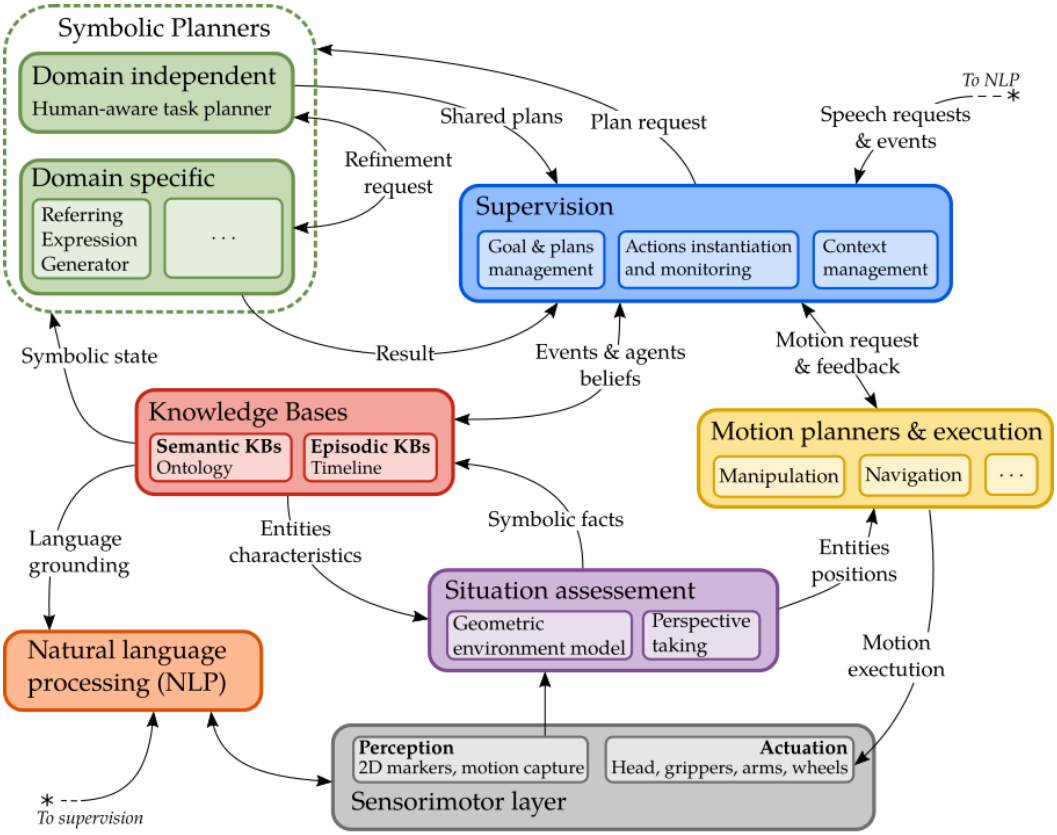


Figure 4.5: An overview of the cognitive architecture developed to handle the Director Task. Each block does not necessarily represent one software component but rather an architectural module (in terms of the features it implements). The arrows represent the type of information exchanged between the modules. This architecture extends the ones presented all along with this thesis.

4.3.1 Storing and reasoning on symbolic statements

The knowledge representation is always a core component of cognitive architectures as organising the knowledge allowing the robot to better understand the environment it evolves in. Moreover, it is on the basis of this knowledge that a robot can communicate with its human partner about the current state of the world and ground the partner's utterance regarding this world state.

Some architectures propagate knowledge all along their components [Hawes 2007], each of them enriching knowledge at each stage before filling it to the next ones. Others consider their knowledge base as an active server, activating perception processes when needed, depending the information we are looking for [Beetz 2018]. For our architecture, we remain on the principle of a central, server-based knowledge base. It is refined into two distinct sub-modules, the semantic knowledge base and the episodic one. The semantic part is in charge of representing the environment elements: the objects' and agents' types, their applicable properties, the descriptions and parameters of the actions, a part of the language model with verbs or pronouns, and their names in natural language. Besides, we also use it to represent the current symbolic world-state (the computed facts) and thus the instantiation of the concepts in terms of physical (e.g., this particular block) or abstract (e.g., this particular action instance) entities. Among these instantiations, we have a part used for the interaction in itself, like the blocks' visual features, and others for the robot programming, like the objects' computer-aided design (CAD) models or tags ids. The episodic knowledge base aims at keeping a trace of the symbolic transitions of the world over time. It is strongly linked to the semantic knowledge base as it allows to semantically interpret these transitions.

The semantic knowledge base is still an ontology managed by the software Ontologenius. The episodic one is in the form of a timeline, managed by the software Mementar[5].

4.3.2 Assessing the world: from geometry to symbolism

The role of the geometrical Situation Assessment module is first to gather different perceptual information and build an internal geometric representation of the world, composed of objects and agents. From this world representation, the module runs reasoning processes to interpret it in terms of symbolic statements between the objects themselves and between the involved agents and the objects. Doing so, the module only builds the robot's representation. However, it does not necessarily reflect what the human partner believes about the world. This is the case with the occluded compartments of the task. If a block is present in a compartment occluded from the human perspective, this block is not visible and thus unknown to the human. Consequently, it should not exist in the human representation of the world. Here is the second role of the Situation Assessment module, estimate the human's perspective and build an estimation of their world representation. It is the first step allowing to implement the theory of mind principles [Baron-Cohen 1985].

To implement this module, we have chosen the Underworld framework [Lemaignan 2018]. Its advantage is to not be monolithic[6]. It works on the principle of a set of worlds, each working at a different granularity and providing

[5] `https://github.com/sarthou/mementar`

[6] It can however be a disadvantage in terms of performance but for research purpose, it allows more flexibility.

specific features, links to create a so-called cascading structure. In the idea, it can be compared to a perception pipeline like [Beetz 2015]. It allows easy reuse of existing modules and makes the core reasoning capabilities independent of the used perception modalities. Even if we choose to use tags for objects detection in this implementation, we could easily switch to machine learning approaches. In the same way, we could use the module with simulations or Virtual Reality systems.

The four worlds we create for the Director Task and their connexions are represented in figure 4.6. At the top (a), we have the perception modalities. For the objects we use AR-tags [Fiala 2005]. For humans, we use a motion capture (mocap) system with helmets equipped with reflectors. For now, only the head is tracked. From each perception input, we create a dedicated world. In these worlds, we can filter the perception data depending on the used system. For the mocap, the data is clean enough. For the AR-tags we apply first a motion filter to discard data acquired when the robot moves. In addition, we apply a field of view (FOV) filter to discard data from the border of the camera because of distortions giving wrong positions even with camera calibration. To know to which object correspond a given tag unique identifiers (UID), the worlds have access to the ontology and can query it to get the UID related to. In the same principle, they can get the objects CAD model. As the output of these worlds, we ensure to have stable data with UID related to the knowledge base.

The world of the middle (b) is the robot's world representation. Information from the perception worlds are merged along with the static elements, like the building walls, and the robot model. From this world, additional perception reasoning processes are applied for the objects that are no more visible in the way of [Milliez 2014b]. If an entity is no more perceived in one of the previous worlds, we first test if it should be in the robot's FOV. If so, the robot should see it. To get an explanation of this absence, we test if another entity could hide it. If not, the object is removed from the world representation. Otherwise, we keep it as we have found an explanation. Once the entities are stabilised, geometric reasoners are applied to them to extract symbolic facts. In the current version of the system, the computed facts are *isOnTopOf*, for an object on top of another with a direct contact, *isInside*, for a block in a compartment, *isVisibleBy*, assessing if an agent could see the object or not from his position, and *isReachableBy*, assessing if an object can be taken by an agent. All these facts are sent to the robot's semantic knowledge base, where reasoners will deduce further facts. For example, if a block is in a compartment, thanks to inverse property *hasInside* the fact that the compartment has the block inside is computed. In the same way, if this compartment is on top of the table, the block inside is computed to be above the table (*isAbove)* thanks to chain axiom.

While the previous world corresponds to the robot's representation, the human partner cannot have the same because of the occluded compartments. The world c) thus aims at estimating the representation of the world from the partner's perspective. From the robot's world, we compute a segmentation image from the human point of view and use it as a filtered perception world. This allows us to instantiate

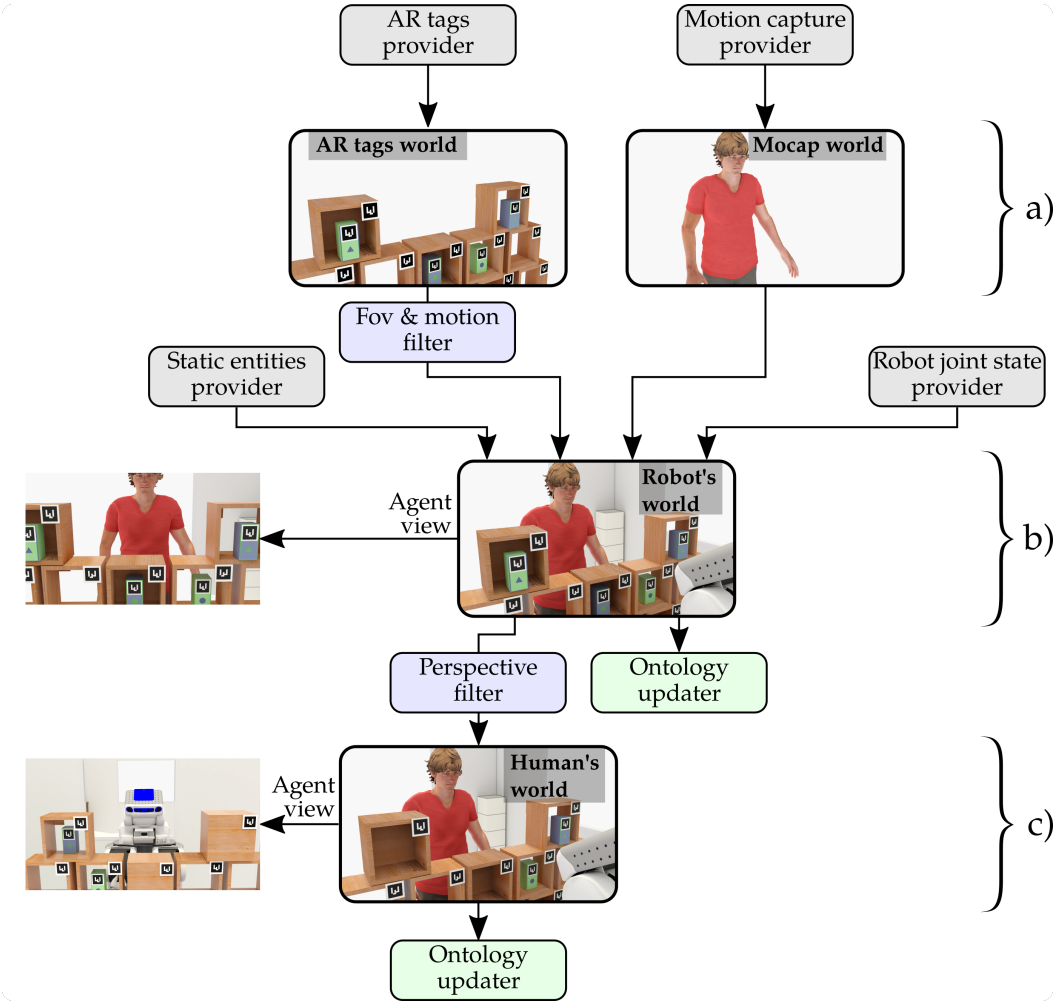


Figure 4.6: The world cascading structure of the geometrical situation assessment system. The two worlds at the top (a) are build from the perception systems and filtered. The world of the middle (b) merges the different perception information and computes symbolic facts on it. The world at the bottom (c) is the estimation of the human world representation and is computed from perspective-taking in the robot's world. Like for the world of the middle, symbolic facts are computed and sent to the semantic knowledge base.

the same world management process we used for the robot but this time for the human. In this way, we emulate their perception capability and geometric reasoning process. Symbolic facts are thus computed and sent to the human's semantic knowledge base. In the world of the bottom (c) on figure 4.6, we can see that the two blocks in the occluded compartments are not present in the human world. Here we make explicit the difference between an object that is unknown and an object that is known but not visible. We could have an interaction where the human goes to see the robot side and the robot would consequently estimate the blocks in the

occluded compartments as known to the human but not visible.

4.3.3 Planning with symbolic facts

The symbolic planners are divided into two categories: the domain-independent one, planning high-level tasks, and the domain-dependant one, specialized in solving precise problems. For the Director Task, the only domain-specific planner used is the Referring Expression Generator.

Where we previously used HATP [Lallement 2014] as task planner, for this task we used its next-generation presented in [Buisan 2021]. In the same way as HATP, the new planner aims at taking into account the human's contribution to planning how to perform a high-level task. To do so, it can generate a shared plan in which parts of the task are assigned to the human partner and other to the robot itself, depending on some criteria. However, the robot's partner is not an agent that the planner can directly control. Indeed, it must sometimes communicate about the plan to inform the human about their next actions. The new planner rather trends at emulating the human decision, action, and reaction processes to generate a shared plan. For the Director Task, emulating the human reaction to a given instruction enables the comparison between multiple blocks order, the communication of higher-level instructions to the human and the balance between multiple communication modalities.

The REG planner has been successfully integrated with the new planner allowing it to estimate the cost and the feasibility of referring communication at task planning level. The initial world state is fetched from the ontology leading to a uniformity of the knowledge among the architecture.

4.3.4 Managing the interaction

The supervision component aims at managing the overall interaction. In this architecture, we use JAHRVIS (Joint Action-based Human-aware supeRVISor) which constitutes the decisional kernel of this cognitive architecture. Like its predecessors, SHARY [Clodic 2009] and its extensions [Fiore 2016, Devin 2016], it is designed for a human-aware robot. It has to not only handle the robot's action execution but also to estimate the human mental state, monitoring his actions, and communicate with him. To handle these features, several processes are needed:

Interaction sessions management: It manages an interaction session that is first refined into tasks, themselves refine into action coming from the task planner. Moreover, it is in charge of the greetings happening at the beginning of an interaction, the goodbyes at the end, and all events and exchanges happening outside tasks (e.g., conversation, goal negotiation) or during a task but not related to it like a human doing a parallel task on its own.

Communication management: Communications are categorized in JAHRVIS either to: give information updating the receiver beliefs; ask a question to update the emitter belief; ask the other agent to perform an action; discuss with dialogue not related to a task or a goal/plan negotiation.

Human management: As the supervision system manages shared plans, it has to make sure the human follows them. Moreover, even if some communications are planned, it also has to make sure that the human has all the knowledge he needs for what he has to perform and if not, it hence acts or communicates through the other processes. To do so, it monitors the human beliefs about the ongoing task and plan.

Task management Even if the human has also the necessary information about the plan, contingencies can happen. The supervision can react and perform a repair thanks to action or communication.

Quality of Interaction management Even if a task is achieved, it could be done more or less efficiently and smoothly. All along an interaction session and a task, the supervision system thus estimates in real-time the Quality of Interaction (QoI) [Mayima 2020]. It measures the human engagement and the effectiveness of collaborative task performance. This information can then be used by the decision-making process to tune dynamically others processes such as the cost of properties for the REG.

4.3.5 Speaking and understanding

The Natural Language Generation is made of two parts, a static one for action verbs and communications to signify a lack of understanding and a dynamic part for the referring expressions. The content is determined by the REG and the linguistic realisation is done on the basis of concepts' labels in the ontology and a simple grammar model to know in which order the adjectives have to be sorted depending on the language.

Natural Language Understanding is more difficult due to the variety of ways the same information can be communicated. Moreover, in a given communication, we have different information. In the Director Task, we have the action to perform and the object on which the action has to be performed. First, we use the Google Speech To Text (STT) API to pass from an audio stream to a string of characters. Even if such technology is now well mastered, mistakes still appear in the transcription[7]. On the string, we perform a first analysis trying to match words and group of words with labels of the ontology. We used sliding windows limited on the length and the fuzzy match technique available with ontologenius. To cover a maximum of possibilities, several action verbs are described as well as synonyms for the concepts.

[7] And this, even more, depending on our English accent and the quality of the microphone used

We also tried to have a good hierarchy in the ontology types for the robot to better catch the concepts depending on the abstraction level used by the human. To refer to the blocks, some only use the terms "object" as they are the only ones involved in the task. At the end of this analysis, we have a list of concepts. Depending on the number of uncaught words (the words unknow in the ontology), we can already know if the understanding is poor or not. On the concept list, we first extract the action verb to know the instructed action (e.g. take, place, remove). The rest of the sentence is analysed thanks to the inverse grammar model for one part but also thanks to the properties ranges and domains. When we said "the red apple", we do not have any word representing the used property[8]. With the analysis of the usable properties linking color to an apple (and thus to a vegetable and so on), we are able to find the corresponding property. The result of this analysis is a SparQL query in the same way such query is used for the NLU. Depending on the number of concepts successfully linked we can estimate the comprehension quality. The SparQL query describing the entity to act upon is then merged with the context of the task and sent to the ontology to find the target entity. In our case, the context would be the same as for the generation meaning that we are speaking about an object being above the table of interaction.

In the case the human gives an accurate description, we should have only one match for the target entity. However, we cannot consider that the human will never do a mistake or that the robot will fully understand the instruction. In this case, we run a REG on all the ambiguous entities. The context of these generations is the SparQL query coming from the understanding process. If we know that we are already speaking of a green block, we do not have to recall it. We fall back into Natural Language Generation and generate sentence like "do you mean the block with a circle or a triangle ?". When the human responds, we use again the SparQL query coming from the first utterance and merge it with the newly understood.

For the Natural Language Understanding part, we could use machine learning approaches based on sequence-to-sequence (seq2seq) models like [Panchbhai 2020]. However, doing so we duplicate the knowledge already existing in the ontology to put it in a neural network. Unless creating a standard of concept identifier, such model should be trained for each used knowledge base in order to be compatible with it and use the same symbols. Having different symbols would lead to failure, having more symbols in the trained model would lead to failure (queries that could not match), and having fewer symbols in the trained model would lead to a lack of understanding. Moreover, in addition, to create the ontology, we would have to create the corresponding training dataset that is a huge amount of work even if artificially augmented dataset creation techniques exist.

Even if our method can be seen as being ha-doc, we ensure uniformity of the knowledge among the architecture. Moreover, it can be easily extended and even dynamically extended during an interaction.

[8]It is often the case of the attributes where relations between entities are more explicit.

4.4 Experiments

The architecture has been successfully implemented on a Pr2 robotic platform. The robot is thus able to play both roles, the director and the receiver. In this section, we comment and analyse a video[9] of two experiments. The only emulated element is the human action recognition to trigger the next actions of the robot and it is the director.

4.4.1 Pr2 as the director

We start this section with a Pr2 in the role of the director (0:21 in the video). The setup is composed of six compartments including two compartments hidden from one side. One of these compartments is hidden from the human (the receiver) and one from the robot (the director). One block has been placed in each compartment. Consequently, only four blocks are known by both the human and the robot. Figure 4.7 is a visualization of the estimated geometric model of the human, maintained by the situation assessment component. Even if a block is present in each compartment, the leftmost one is not present in the estimation of the human's world. This absence comes from the fact that the human can not see what is in the compartment and thus can not know this block.

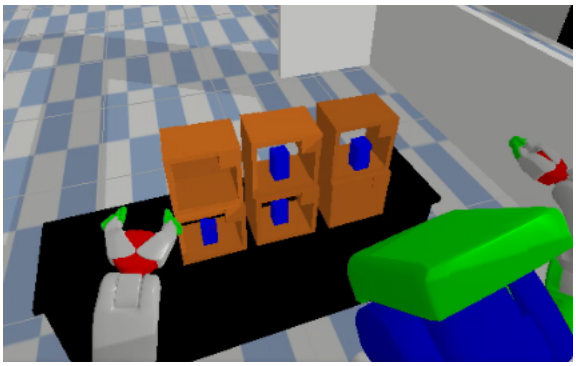


Figure 4.7: A visualization of the human's estimated goemetric world. Even if a block is present in each compartment, the left most one is not present in this worls since the human can not see this block.

Figure 4.8 represents the entire interaction when the robot is the director. At the initial state, four blocks are visible from both agents. Describing them with all their visual features, they are:

- A blue block with a blue border and a green triangle
- A blue block with a blue border and a green circle
- A blue block with a green border and a blue triangle

[9] `https://youtu.be/jtSyZeqBkp0`

- A green block with a green border and a blue circle

Thanks to the estimation of the communication cost at task planning using the results of the REG, the robot is able to find the optimal sequence of blocks to intrust. The overall communication is thus minimized and the RE is unambiguous in each situation. In the initial state (a to b), the robot asks for the green block as only one of the visible blocks is green. Since the green block has a circle on it, removing it, only one of the remaining blocks has a circle on it. The robot can thus use this feature to refer to the next block (b to c).

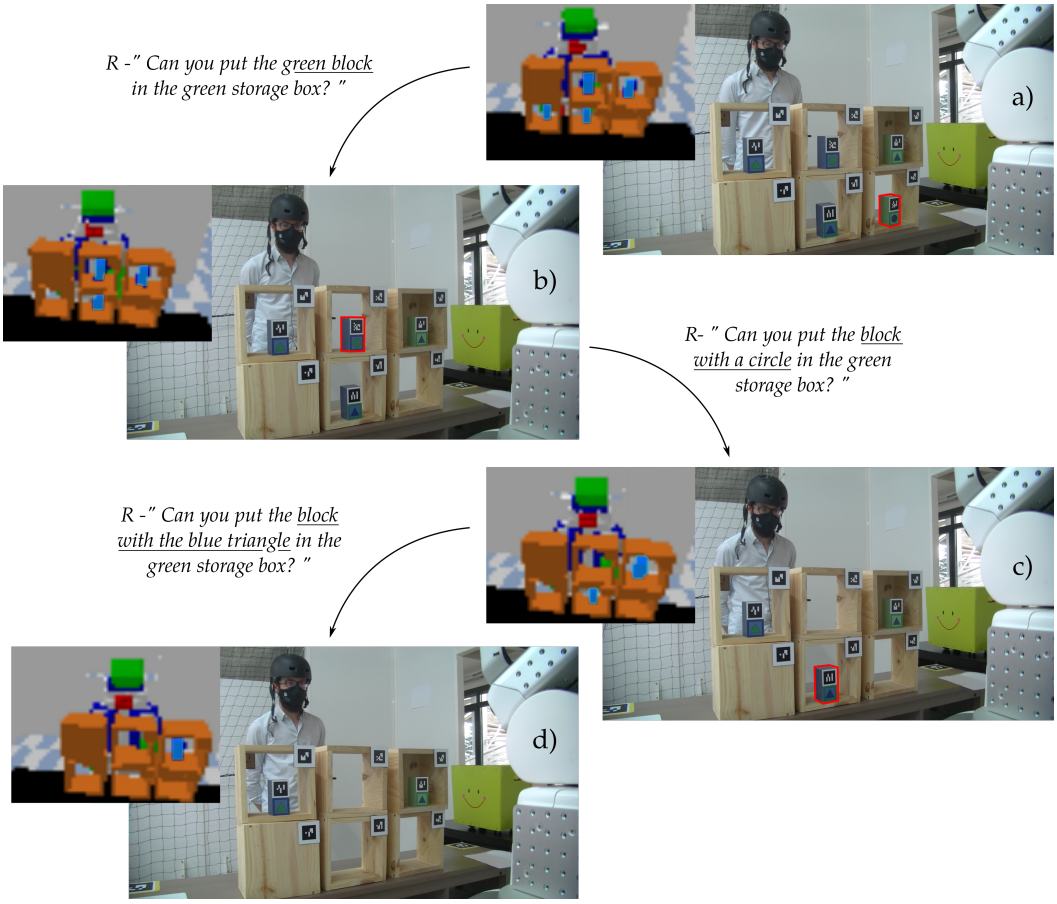


Figure 4.8: The director task handled by an autonomous PR2 robot in the role of the director. Each picture represents a step toward the achievement of the task. The estimated human perspective is displayed in the top left-hand corner of each picture. On top of the arrows leading to a new state are the sentences said by the robot to the human. The block outlined in red are the blocks referred to at each step.

4.4.2 Pr2 as the receiver

While in its previous role the robot just had to instruct the human, when the robot is the receiver (1:33 in the video) more reasoning is needed. A retranscription of part of the interaction is represented in figure 4.9. In the initial state, the same four blocks as previously are visible by both the agents. The robot is able to understand three actions: take, drop, and remove. The latter action is a combination of the two others.

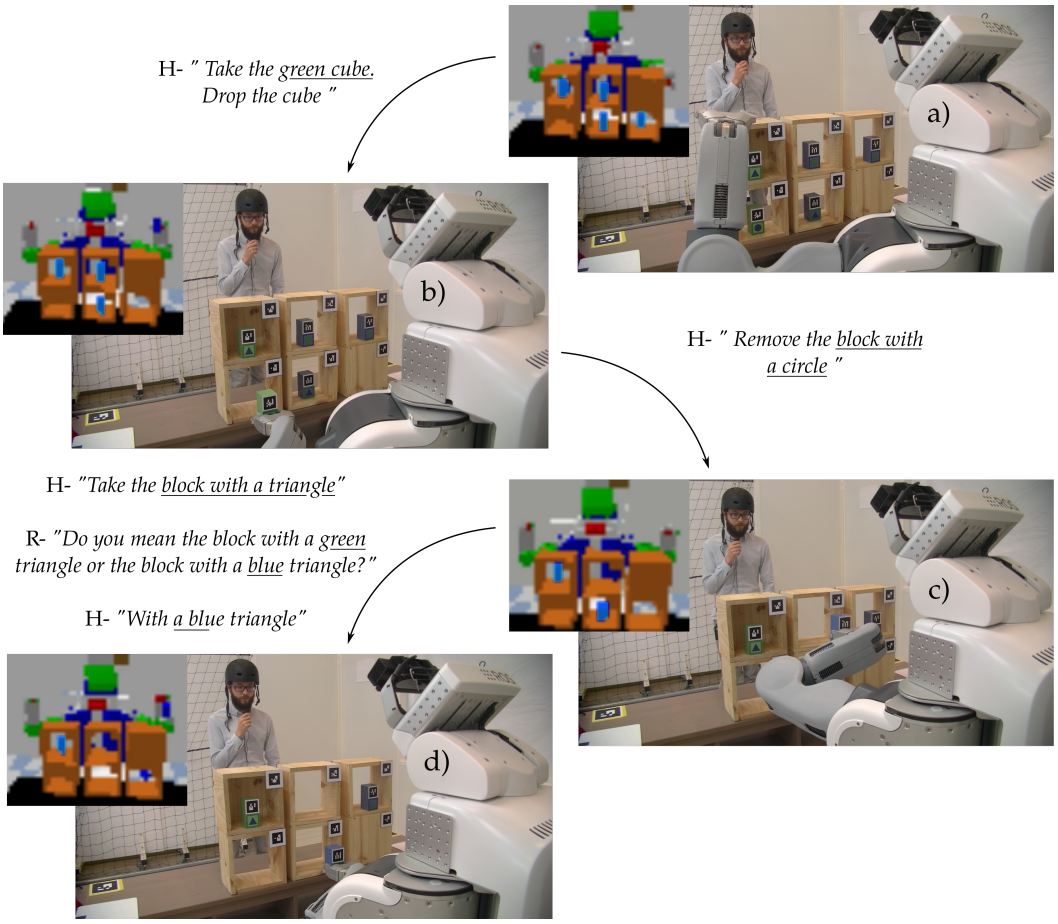


Figure 4.9: The director task handled by an autonomous PR2 robot in the role of the receiver. Each picture represents a step toward the achievement of the task. The estimated human perspective is displayed in the top left-hand corner of each picture. On top of the arrows leading to a new state are the sentences said by the human to the human and for the last situation the refinement query from the robot to the human.

For the first block (a to b on the figure), the human intruct the robot for the green block. The natural language understanding module return the SparQL query:

(?0, isA, Block), (?0, hasColor, green)

Since the robot assumes the human to speak about objects on the table, the understood query is merged with another one representing the context of the task: (?0, isAbove, table_1). Querying the human estimated ontology with the merged query, only one entity match. There is no ambiguity in human instruction. The robot takes the instructed block then drop it. If the query was applied to the robot ontology, two blocks would have matched since the block unknown by the human is also green. It goes the same for, the second instruction. There is no ambiguity. The SparQL query related to this second block is:

(?0, isA, Block), (?0, hasFigure, ?1), (?1, isA, Circle)

The third instruction given by the human as the director is the most interesting for us. The human asks for *"The block with a triangle"*. However, the speech to text returns *"take is about to whip a triangle"*. With this sentence, the NLU module can only extract two known concepts being "take" and "triangle". Due to the limited amount of word understood, it does not try to generate a SparQL query. The robot thus informs the human about its incapacity and repeat the heard sentence as a back loop for the human. At the second try, the sentence is understood and gives the query:

(?0, isA, Block), (?0, hasFigure, ?1), (?1, isA, Triangle)

However, matching this query to the human's estimated ontology, we get two results. Once again, matching it to the robot's ontology would give three results but the third one is not visible from the human. Since all the concepts of the sentence have been understood and linked together to create the query, the human should have made a mistake, providing an ambiguous referring expression.

To be proactive, we want the robot to ask precision about the block to take by proposing visual features to distinguish them. To do so, we use the REG algorithm on each ambiguous block. As a context for the REG, we pass the previously merged SparQL query. It represents what has already been understood by the robot. In the current situation, the robot thus performs two REG and their results are used to generate the disambiguation sentence:

"Do you mean the block with a green triangle or the block with a blue triangle?"

When the human responds, for sure it does not generate a complete description f the block to be taken. It rather answers the question. The query extracted from his answer is thus combined with the previously understood one in case some information is missing. Matching this last query to the human's estimated ontology, the robot finally get the block to remove.

With this latter case, we saw how the robot can react to a human's mistake and use the REG to help the progress of the task, even if it is the receiver.

4.5 Open challenges for the community

So far, we have described the main abilities a robot has to be endowed with to perform the Director Task. Then, we have proposed a cognitive robot architecture handling the Director Task in its simplest form, both for the director and receiver roles. However, we have only tackled the normals cases that the task offers. In this section, we now present some open challenges that we have identified around the task. In addition, since we see that the environment of the task can be controlled, we also propose some user studies to investigate the ways of sharing information.

4.5.1 Challenges to take up

The components or abilities related to each challenge are reported in the following table. The list of challenges is not exhaustive. Moreover, even if some challenges have already been mentioned among the presentation of the components, they are here reported as requiring finer and more generic management.

Challenged abilities / components	Challenges
Perspective-taking	1
Communication	4, 6, 7
Task planning	2, 3, 4
Reference generation	4, 5, 8
Contingencies handling	1, 2, 3, 4

1. **Finer contingency analysis:** In this task, failures can easily arise due to the high ambiguity between the blocks and the difference of perspective. Such failures have to be handle by the robot and to do so their origin have to be understood to react to them in an appropriate meaner. In the case the human as the receiver does not take the instructed block, the failure can have different origins. First, it could come from a perspective not taken into account. However, this lack of perspective-taking can be assigned either to the director or the receiver. Another origin can be a description not clear enough or correct but too complex. Finally, it can just be an error of inattention. Each of these origins has to be handle in a different way.

2. **Handling contingencies as errors:** When the receiver takes another block than the one instructed, has to fix the error through communication and negotiation. First, the wrong block has to be put back in its original compartment. Then, the robot has to adapt its original instruction to make it clearer and improve the chances to have the receiver taking the right one.

3. **Not handling contingencies as errors:** When the receiver takes the wrong block, even if it is the instructed one, it can however be part of the goal. In this case, the robot not necessarily has to repair the plan, asking the human to put it back as no order is required for the task. It can thus re-plan or

re-instruct the human for the same block without further information. It may also mention to the receiver for the mistake and explain that it does not matter because this one is also part of the goal. Rather than re-planning, the robot could use a conditional plan, anticipating possible confusions, and adapt according to the human's actions.

4. **Adapting to recurrent failures:** In case of recurrent failures by the partner or degradation of the Quality of interaction with numbers of latencies, the robot could try to analyse the origin of the problems and determine if a common point exists. If so, it can adapt itself to increase the QoI and reduce the failures. For example, if the partner is found to have difficulties with certain visual features, the robot can react through properties' cost adaptation. If the partner still consider the removed blocks, it can react through communication context adaptation.

5. **Allowing spatial references:** As explained in section in the origins of the task, the Director Task is originally a task to test referential communication. Even if the present version asks the participants to not use spatial reference, this rule could be relaxed to study perspective-corrected spatial Referring Expression Generation.

6. **Understanding the human instructions:** In the current implemented version, the robot can only understand a limited vocabulary and restricted to the context of the task. In this way, the robot only understands descriptions of blocks. In a more natural interaction, humans could use a richer vocabulary, give a single instruction in multiple steps, or have communications not directly linked to the task. During tests for designing the task, it was common to have instructions like "take the block with a ... triangle. No, rather the one with a green border". Such complex communications where the director corrects his explanations should have to be managed by the robot.

7. **Introducing code words:** As presented through the design of the used material, the visual features on the blocks have been chosen in a way to allow the visualisation of landscapes on them, with a little imagination. Considering multiple tasks with the same robot and human, alternating the roles if needed, the introduction of coded words could be interesting to reduce the communication complexity and thus the overall efficiency. The robot could thus try to negotiate some coded words. Once introduced, it would also have to remember them and understand them as being part of a description.

8. **Communicating about multiple blocks:** With the currently implemented system, the director only instruct one block at a time. It can either be through a reference matching all of them, like "Take all the blocks with a triangle on them", or multiple descriptions in a raw. The latter method could bring different kinds of communications such as "I do not remember the instruction for the last block" when the human is the receiver. For the first method, when

the robot is the receiver, it would also be a different kind of instructions to interpret.

4.5.2 User studies to perform

Some robot behaviours, mainly about the referring expression generation, have been designed with regard to the current literature. However, the Director Task could be used to refine them thanks to user studies. More than providing a controlled task and environment, this task has the advantage to hide the real goal of the study. From the participant point of view, the goal is to remove blocks from compartments. The goal of the study can be focused on other aspects and could help the community in the design of architectures applied to more realistic scenarios.

Currently, the references to the blocks are made in such a way as to minimize the number of visual features used while staying discriminative. Such implementation fit Grice's Maxim of Quantity [Grice 1975]. However, due to all the cognitive mechanisms to use in this task (e.g., perspective-taking) and the high ambiguity among the blocks, evaluating such behaviour compared to a full explanation could be interesting. Indeed, giving a reference with more information than needed would ensure to not match blocks being only visible by the receiver, which could help them to select the right block. In a way, it could allow to not use perspective-taking at the cost of complex communications.

During the material presentation, we have introduced a special compartment equipped with a wire mesh. Because a block in such a compartment is visible from the receiver but not accessible, referring to a block matching also this one could disturb the receiver. We could expect such a situation to require a higher cognitive load to determine the right block to take. Such behavior could also be interesting to evaluate as even if the human receiver is able to take the right block it could also decrease the Quality of Interaction. In the same way, a block previously visible by the receiver and that the director moves in a hidden compartment could disturb the receiver to interpret a description.

Chapter 5

A robot in a storage room

Conclusion

On the Human Agent Interaction Guidelines

Limitations and Future Work

Notes

Appendix A

Appendix 1

As the metrics are aggregated to compute the QoI, their values need to be on the same scale. In order to do this, we use scaling functions rescaling metrics into a range of $[-1, 1]$, as the QoI bounds. As all the metrics does not have the same properties, they have to be scaled by using different functions. The two properties to check to choose which function to apply to which metric are the following ones:

- does the metric already have a bounded value ?
- what value of the metric should make the QoI decrease, increase or remain the same ?

Therefore, we designed three functions to be used with metrics having bounded values and three functions for metrics that do not have upper bounds. Then, among these two sets of functions, it is possible to choose the one to use according to the positive, neutral or negative impact a value should have on the QoI.

A.1 Scaling of bounded metrics: Min-Max Normalization

We defined three min-max normalization functions, illustrated in Fig. A.1. They were designed to be used for metrics whose values belong to a bounded set, i.e., metrics for which the minimum and maximum values are known. The first function is to apply in cases for which a measure approaching the bound value b_1 has a negative impact on the quality evaluation whereas a measure approaching b_2 has a positive one. It allows to scale a measure x between -1 and 1:

$$n_1(x) = 2 * \frac{x - b_1}{b_2 - b_1} - 1 \tag{A.1}$$

The second function is intended to be applied in cases for which a measure approaching the bound value b_1 has a neutral impact on the quality evaluation whereas a measure approaching b_2 has a positive one. It allows to scale a measure x between 0 and 1:

$$n_2(x) = \frac{x - b_1}{b_2 - b_1} \tag{A.2}$$

Finally, the last function is to apply in cases for which a measure approaching the bound value b_1 has an negative impact on the quality evaluation whereas a measure

approaching b_2 has a neutral one. It allows to scale a measure x between -1 and 0:

$$n_3(x) = \frac{x - b_2}{b_2 - b_1} \tag{A.3}$$

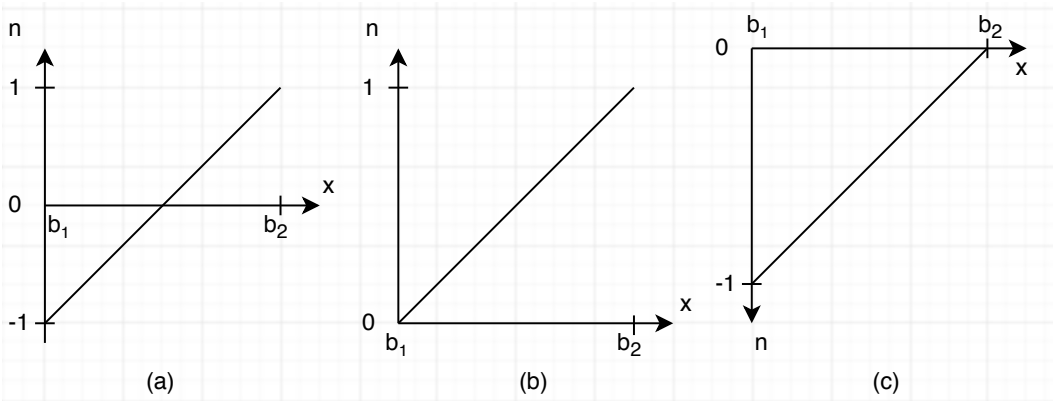


Figure A.1: (a), (b) and (c) respectively represent the min-max normalization functions (A.1), (A.2) and (A.3)

A.2 Scaling of unbounded metrics: Sigmoid Normalization

We defined three sigmoid-like functions to scale and squash values of metrics without an upper bound. As for the min-max normalization, there is one function to scale the metrics values between -1 and 1, another one to scale between 0 and 1 and the last one to scale between -1 and 0.

The first function allows to scale between -1 and 1 the values of a metric, for a metric whose values are between 0 and $+\infty$ (e.g. a duration whose final value is unknown during the execution). The function is defined as:

$$s_1(x) = 1 - 2\exp\left(-\ln(2)\left(\frac{x}{th}\right)^k\right), x > 0 \tag{A.4}$$

with $s_1(x) \in [-1, 1]$, th the value of the sigmoid's midpoint (i.e., $s_1(th) = 0$) and, k setting the shape of the function curve. k and th values are set off-line by the designer and they allow to define the shape of the metric scaling.

The second function is designed for metric which cannot have a negative impact on the QoI as it scales the value between 0 and 1 (and with $x \in [0, +\infty]$ as well):

$$s_2(x) = 1 - \exp\left(-\ln(2)\left(\frac{x}{th}\right)^k\right), x > 0 \tag{A.5}$$

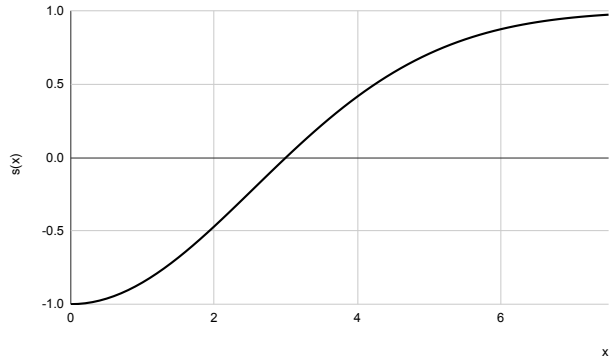
with $s_2(x) \in [0, 1]$, th the value of the sigmoid's midpoint (i.e., $s_2(th) = 0.5$) and, k setting the shape of the function curve.

The third function is designed for metric which cannot have a positive impact on the QoI as it scales the value between -1 and 0 (and with $x \in [0, +\infty]$ as well):

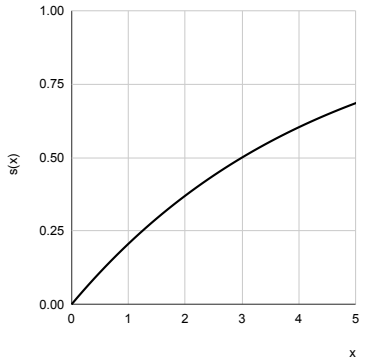
$$s_3(x) = -1 + \exp\left(-\ln(2)\left(\frac{x}{th}\right)^k\right), x > 0 \tag{A.6}$$

with $s_3(x) \in [-1, 0]$, th the value of the sigmoid's midpoint (i.e., $s_3(th) = -0.5$) and, k setting the shape of the function curve.

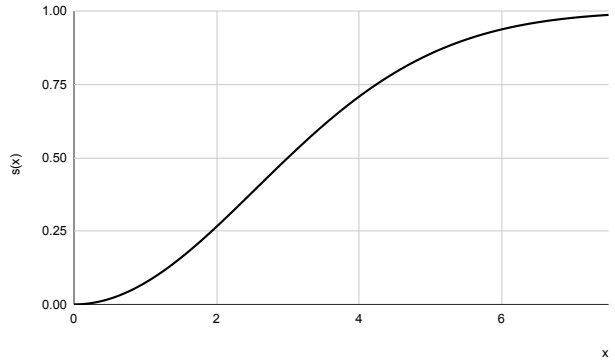
The functions $s_1(x)$ and $s_2(x)$ are illustrated in Fig. A.2 with four examples.



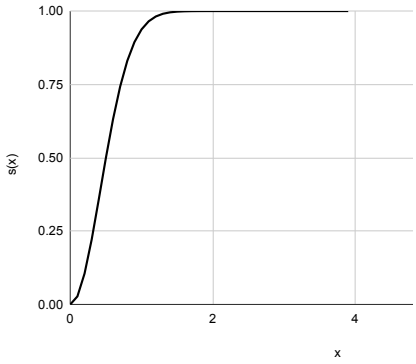
(a) Plot of $s_1(x)$ with $th = 3$ and $k = 2$



(b) Plot of $s_2(x)$ with $th = 3$ and $k = 1$



(c) Plot of $s_2(x)$ with $th = 3$ and $k = 2$



(d) Plot of $s_2(x)$ with $th = 0.5$ and $k = 2$

Figure A.2: Plots of the sigmoid-like functions $s_1(x)$ and $s_2(x)$ with different parameters values

Bibliography

[Allen 2003] Gary L. Allen. *Gestures Accompanying Verbal Route Directions: Do They Point to a New Avenue for Examining Spatial Representations?* Spatial Cognition & Computation, vol. 3, no. 4, pages 259–268, 2003. (Cited in page 42.)

[Amici 2015] Federica Amici and Lucas M Bietti. *Coordination, collaboration and cooperation: Interdisciplinary perspectives.* Interaction Studies, vol. 16, no. 3, pages vii–xii, 2015. (Cited in page 12.)

[Anzalone 2015] Salvatore M. Anzalone, Sofiane Boucenna, Serena Ivaldi and Mohamed Chetouani. *Evaluating the Engagement with Social Robots.* International Journal of Social Robotics, vol. vol. 7, no. 4, pages pp. 465–478, Aug 2015. (Cited in pages 25 and 76.)

[Argyle 1973] Michael Argyle. Social Interaction. Transaction Publishers, 1973. (Cited in page 4.)

[Baraglia 2017] Jimmy Baraglia, Maya Cakmak, Yukie Nagai, Rajesh P. N. Rao and Minoru Asada. *Efficient human-robot collaboration: When should a robot take initiative?* International Journal of Robotics Research, vol. vol. 36, no. 5-7, pages pp. 563–579, 2017. (Cited in page 23.)

[Baron-Cohen 1985] Simon Baron-Cohen, Alan M Leslie and Uta Frith. *Does the autistic child have a "theory of mind"?* Cognition, vol. 21, no. 1, pages 37–46, 1985. (Cited in page 90.)

[Bauer 2008] Andrea Bauer, Dirk Wollherr and Martin Buss. *Human–robot Collaboration: A Survey.* International Journal of Humanoid Robotics, vol. vol. 5, no. 01, pages pp. 47–66, 2008. (Cited in page 22.)

[Bauer 2009] Andrea Bauer, Klaas Klasing, Tingting Xu, Stefan Sosnowski, Georgios Lidoris, Quirin Muhlbauer, Tinguang Zhang, Florian Rohrmuller, Dirk Wollherr, Kolja Kuhnlenz*et al. The autonomous city explorer project.* In IEEE International Conference on Robotics and Automation (ICRA), pages 1595–1596. IEEE, 2009. (Cited in page 38.)

[Becchio 2010] Cristina Becchio, Luisa Sartori and Umberto Castiello. *Toward you: The social side of actions.* Current Directions in Psychological Science, vol. 19, no. 3, pages 183–188, 2010. (Cited in page 12.)

[Beetz 2015] Michael Beetz, Ferenc Bálint-Benczédi, Nico Blodow, Daniel Nyga, Thiemo Wiedemeyer and Zoltán-Csaba Marton. *Robosherlock: Unstructured information processing for robot perception.* In IEEE International Conference on Robotics and Automation (ICRA), pages 1549–1556, 2015. (Cited in page 91.)

[Beetz 2018] Michael Beetz, Daniel Beßler, Andrei Haidu, Mihai Pomarlan, Asil Kaan Bozcuoğlu and Georg Bartels. *Know rob 2.0—a 2nd generation knowledge processing framework for cognition-enabled robotic agents.* In IEEE International Conference on Robotics and Automation (ICRA), pages 512–519. IEEE, 2018. (Cited in page 90.)

[Bekele 2014] Esubalew Bekele and Nilanjan Sarkar. *Psychophysiological Feedback for Adaptive Human–Robot Interaction (HRI).* In Stephen H. Fairclough and Kiel Gilleade, editors, Advances in Physiological Computing, pages pp. 141–167. Springer London, 2014. (Cited in page 26.)

[Belhassein 2017] Kathleen Belhassein, Aurélie Clodic, Hélene Cochet, Marketta Niemelä, Päivi Heikkilä, Hanna Lammi and Antti Tammela. *Human-human guidance study.* Technical Report, 2017. (Cited in pages 41 and 42.)

[Bensch. 2017] S. Bensch., A. Jevtić. and T. Hellström. *On Interaction Quality in Human-Robot Interaction.* In Proceedings of the 9th International Conference on Agents and Artificial Intelligence (ICAART), pages pp. 182–189, 2017. (Cited in page 26.)

[Bethel 2010] Cindy L. Bethel and Robin R. Murphy. *Review of Human Studies Methods in HRI and Recommendations.* International Journal of Social Robotics, vol. vol. 2, no. 4, pages pp. 347–359, Dec 2010. (Cited in page 25.)

[Bohus 2014] Dan Bohus, Chit W Saw and Eric Horvitz. *Directions Robot: In-the-Wild Experiences and Lessons Learned.* In Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), page 8, 2014. (Cited in page 39.)

[Brass 2001] Marcel Brass, Harold Bekkering and Wolfgang Prinz. *Movement observation affects movement execution in a simple response task.* Acta psychologica, vol. 106, no. 1-2, pages 3–22, 2001. (Cited in page 15.)

[Bratman 1987] Michael Bratman*et al.* Intention, plans, and practical reason, volume 10. Harvard University Press Cambridge, MA, 1987. (Cited in page 16.)

[Bratman 1988] Michael E Bratman, David J Israel and Martha E Pollack. *Plans and resource-bounded practical reasoning.* Computational intelligence, vol. 4, no. 3, pages 349–355, 1988. (Cited in page 16.)

[Bratman 2013] Michael E Bratman. Shared agency: A planning theory of acting together. Oxford University Press, 2013. (Cited in page 15.)

[Brawer 2018] Jake Brawer, Olivier Mangin, Alessandro Roncone, Sarah Widder and Brian Scassellati. *Situated Human–Robot Collaboration: predicting intent from grounded natural language.* In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 827–833. IEEE, 2018. (Cited in page 81.)

[Buisan 2021] Guilhem Buisan and Rachid Alami. *A Human-Aware Task Planner Explicitly Reasoning About Human and Robot Decision, Action and Reaction.* In Companion of the ACM/IEEE International Conference on Human-Robot Interaction, pages 544–548, 2021. (Cited in page 93.)

[Burgard 1999] Wolfram Burgard, Armin B Cremers, Dieter Fox, Dirk Hähnel, Gerhard Lakemeyer, Dirk Schulz, Walter Steiner and Sebastian Thrun. *The museum tour-guide robot RHINO.* In Autonome Mobile Systeme 1998, pages 245–254. Springer, 1999. (Cited in page 38.)

[Byrne 1997] Michael D. Byrne and Susan Bovair. *A Working Memory Model of a Common Procedural Error.* Cognitive Science, vol. 21, no. 1, pages 31–61, 1997. (Cited in page 60.)

[Caniot 2020] Maxime Caniot, Vincent Bonnet, Maxime Busy, Thierry Labaye, Michel Besombes, Sebastien Courtois and Edouard Lagrue. *Adapted Pepper.* Technical Report, SoftBank Robotics Europe, 2020. (Cited in page 63.)

[Carpenter 2009] Malinda Carpenter. *Just how joint is joint action in infancy?* Topics in Cognitive Science, vol. 1, no. 2, pages 380–392, 2009. (Cited in page 12.)

[Castro 2020] Víctor Fernández Castro and Elisabeth Pacherie. *Joint actions, commitments and the need to belong.* Synthese, pages 1–30, 2020. (Cited in page 13.)

[Chalmeau 1995] Raphaël Chalmeau and Alain Gallo. *La coopération chez les primates.* L'Année psychologique, vol. 95, no. 1, pages 119–130, 1995. (Cited in page 12.)

[Chen 2017] Yingfeng Chen, Feng Wu, Wei Shuai and Xiaoping Chen. *Robots serve humans in public places—KeJia robot as a shopping assistant.* International Journal of Advanced Robotic Systems, vol. 14, no. 3, pages 1–20, 2017. (Cited in page 38.)

[Chong 2007] Hui-Qing Chong, Ah-Hwee Tan and Gee-Wah Ng. *Integrated cognitive architectures: a survey.* Artificial Intelligence Review, vol. 28, no. 2, pages 103–130, 2007. (Cited in page 16.)

[Clark 1996] Herbert H Clark. Using language. Cambridge university press, 1996. (Cited in pages 12, 13, and 15.)

[Clark 2006] Herbert H Clark. *Social actions, social commitments.* In Roots of human sociality, pages 126–150. Routledge, 2006. (Cited in page 13.)

[Clodic 2006] Aurélie Clodic, Sara Fleury, Rachid Alami, Raja Chatila, Gérard Bailly, Ludovic Brethes, Maxime Cottret, Patrick Danes, Xavier Dollat,

Frédéric Eliseiet al. *Rackham: An interactive robot-guide.* In IEEE International Symposium on Robot and Human Interactive Communication (ROMAN), pages 502–509. IEEE, 2006. (Cited in page 38.)

[Clodic 2009] Aurélie Clodic, Hung Cao, Samir Alili, Vincent Montreuil, Rachid Alami and Raja Chatila. *Shary: a supervision system adapted to human-robot interaction.* In Experimental robotics, pages 229–238. Springer, 2009. (Cited in page 93.)

[Clodic 2017] Aurélie Clodic, Elisabeth Pacherie, Rachid Alami and Raja Chatila. *Key Elements for Human Robot Joint Action.* In Sociality and Normativity for RobotsPhilosophical Inquiries into Human-Robot Interactions, Studies in the Philosophy of Sociality, pages 159–177. Springer, 2017. (Cited in page 13.)

[Cohen 1990] Philip R Cohen and Hector J Levesque. *Intention is choice with commitment.* Artificial intelligence, vol. 42, no. 2-3, pages 213–261, 1990. (Cited in page 17.)

[Cohen 1991] Philip R Cohen and Hector J Levesque. *Teamwork.* Nous, vol. 25, no. 4, pages 487–512, 1991. (Cited in page 12.)

[Curioni 2019] Arianna Curioni, Gunther Knoblich, Natalie Sebanz, A Goswami and P Vadakkepat. *Joint action in humans: A model for human-robot interactions.* Humanoid Robotics: A Reference, eds Goswami A, Vadakkepat P (Springer, Dordrecht, The Netherlands), pages 2149–2167, 2019. (Cited in pages 15, 39, and 40.)

[Dautenhahn 2002] Kerstin Dautenhahn, Bernard Ogden and Tom Quick. *From embodied to socially embedded agents–implications for interaction-aware robots.* Cognitive Systems Research, vol. 3, no. 3, pages 397–428, 2002. (Cited in pages 9 and 10.)

[Devin 2016] Sandra Devin and Rachid Alami. *An Implemented Theory of Mind to Improve Human-Robot Shared Plans Execution.* In The Eleventh ACM/IEEE International Conference on Human Robot Interaction (HRI), pages pp. 319–326, Christchurch, New Zealand, 2016. (Cited in pages 22 and 93.)

[Dumontheil 2010] Iroise Dumontheil, Ian A Apperly and Sarah-Jayne Blakemore. *Online usage of theory of mind continues to develop in late adolescence.* Developmental science, vol. 13, no. 2, pages 331–338, 2010. (Cited in page 83.)

[Enriquez 2017] Eugène Enriquez, Levy André and Jacqueline Barus-Michel. Vocabulaire de psychosociologie. Érès, 2017. (Cited in page 5.)

[Fan 2017] J. Fan, D. Bian, Z. Zheng, L. Beuscher, P. A. Newhouse, L. C. Mion and N. Sarkar. *A Robotic Coach Architecture for Elder Care (ROCARE) Based*

on Multi-User Engagement Models. IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. vol. 25, no. 8, pages pp. 1153–1163, 2017. (Cited in pages 29 and 30.)

[Fiala 2005] Mark Fiala. *ARTag, a fiducial marker system using digital techniques.* In IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 2, pages 590–596. IEEE, 2005. (Cited in page 91.)

[Fiebich 2013] Anika Fiebich and Shaun Gallagher. *Joint attention in joint action.* Philosophical Psychology, vol. 26, no. 4, pages 571–587, 2013. (Cited in page 12.)

[Fiore 2016] Michelangelo Fiore, Aurélie Clodic and Rachid Alami. *On planning and task achievement modalities for human-robot collaboration.* In Experimental Robotics, pages 293–306. Springer, 2016. (Cited in page 93.)

[Foster 2019a] Mary Ellen Foster, Bart Craenen, Amol Deshmukh, Oliver Lemon, Emanuele Bastianelli, Christian Dondrup and et al. *MuMMER: Socially Intelligent Human-Robot Interaction in Public Spaces.* In AAAI 2019 Fall Symposium Series, Arlington, United States, November 2019. (Cited in page 64.)

[Foster 2019b] 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.* In Artificial Intelligence for Human-Robot Interaction Symposium (AI-HRI), Arlington, VA, United States, 2019. AAAI Fall Symposium Series 2019. (Cited in page 46.)

[Gaschler 2012] Andre Gaschler, Kerstin Huth, Manuel Giuliani, Ingmar Kessler, Jan de Ruiter and Alois Knoll. *Modelling state of interaction from head poses for social human-robot interaction.* In Proceedings of the Gaze in Human-Robot Interaction Workshop held at the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2012), 2012. (Cited in pages 7 and 9.)

[Georgeff 1989] Michael Georgeff and Felix Ingrand. *Decision-making in an embedded reasoning system.* In International Joint Conference on Artificial Intelligence, 1989. (Cited in page 17.)

[Georgeff 1991] M Georgeff and A Rao. *Modeling rational agents within a BDI-architecture.* In Proc. 2nd Int. Conf. on Knowledge Representation and Reasoning (KR'91). Morgan Kaufmann, pages 473–484. of, 1991. (Cited in page 17.)

[Ghallab 1998] Malik Ghallab, Craig Knoblock, David Wilkins, Anthony Barrett, Dave Christianson, Marc Friedman and et al. *PDDL - The Planning Domain Definition Language*, 08 1998. (Cited in page 22.)

[Ghallab 2016] Malik Ghallab, Dana S. Nau and Paolo Traverso. Automated planning and acting. Cambridge University Press, 2016. (Cited in page 20.)

[Gockley 2005] Rachel Gockley, Allison Bruce, Jodi Forlizzi, Marek Michalowski, Anne Mundell, Stephanie Rosenthal, Brennan Sellner, Reid Simmons, Kevin Snipes, Alan C Schultz*et al. Designing robots for long-term social interaction.* In 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 1338–1343. IEEE, 2005. (Cited in page 8.)

[Goffman 1967] Erving Goffman. Interaction ritual: Essays on face-to-face interaction. Aldine, 1967. (Cited in page 4.)

[Goffman 1983] Erving Goffman. *The interaction order: American Sociological Association, 1982 presidential address.* American sociological review, vol. 48, no. 1, pages 1–17, 1983. (Cited in page 10.)

[Gräfenhain 2013] Maria Gräfenhain, Malinda Carpenter and Michael Tomasello. *Three-year-olds' understanding of the consequences of joint commitments.* PLoS One, vol. 8, no. 9, page e73039, 2013. (Cited in page 12.)

[Grice 1975] Herbert P Grice. *Logic and conversation.* In Speech acts, pages 41–58. Brill, 1975. (Cited in page 102.)

[Gross 2009] H-M Gross, H Boehme, Ch Schroeter, Steffen Müller, Alexander König, Erik Einhorn, Ch Martin, Matthias Merten and Andreas Bley. *TOOMAS: interactive shopping guide robots in everyday use-final implementation and experiences from long-term field trials.* In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 2005–2012. IEEE, 2009. (Cited in page 38.)

[Grosz 1996] Barbara J. Grosz and Sarit Kraus. *Collaborative plans for complex group action.* Artificial Intelligence, vol. vol. 86, no. 2, pages pp. 269–357, 1996. (Cited in page 22.)

[Happé 1994] Francesca GE Happé. *An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults.* Journal of autism and Developmental disorders, vol. 24, no. 2, pages 129–154, 1994. (Cited in page 83.)

[Hawes 2007] Nick Hawes, Michael Zillich and Jeremy Wyatt. *BALT & CAST: Middleware for cognitive robotics.* In IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pages 998–1003. IEEE, 2007. (Cited in page 90.)

[Heesen 2017] Raphaela Heesen, Emilie Genty, Federico Rossano, Klaus Zuberbühler and Adrian Bangerter. *Social play as joint action: A framework to study the evolution of shared intentionality as an interactional achievement.* Learning & behavior, vol. 45, no. 4, pages 390–405, 2017. (Cited in page 13.)

[Heikkilä 2018] Päivi Heikkilä, Hanna Lammi and Kathleen Belhassein. *Where Can I Find a Pharmacy? -Human-Driven Design of a Service Robot's Guidance Behaviour.* In Workshop on Public Space Human-Robot Interaction (PubRob) as part of the International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI), pages 1–2, 2018. (Cited in page 42.)

[Heikkilä 2019] Päivi Heikkilä, Hanna Lammi, Marketta Niemelä, Kathleen Belhassein, Guillaume Sarthou, Antti Tammela, Aurélie Clodic and Rachid Alami. *Should a robot guide like a human? A qualitative four-phase study of a shopping mall robot.* In International Conference on Social Robotics (ICSR), pages 548–557. Springer, 2019. (Cited in page 42.)

[Hiatt 2017] Laura M. Hiatt, Cody Narber, Esube Bekele, Sangeet S. Khemlani and J. Gregory Trafton. *Human modeling for human-robot collaboration.* International Journal of Robotics Research, vol. vol. 36, no. 5-7, pages pp. 580–596, 2017. (Cited in page 22.)

[Hoffman 2007] G. Hoffman and C. Breazeal. *Cost-Based Anticipatory Action Selection for Human–Robot Fluency.* IEEE Transactions on Robotics, vol. vol. 23, no. 5, pages pp. 952–961, 2007. (Cited in page 23.)

[Hoffman 2019] G. Hoffman. *Evaluating Fluency in Human–Robot Collaboration.* IEEE Transactions on Human-Machine Systems, vol. 49, no. 3, pages pp. 209–218, June 2019. (Cited in pages 25 and 76.)

[Ingrand 2017] Félix Ingrand and Malik Ghallab. *Deliberation for autonomous robots: A survey.* Artificial Intelligence, vol. vol. 247, pages pp. 10–44, June 2017. (Cited in page 20.)

[Iocchi 2015] Luca Iocchi, Maria Teresa Lázaro, Laurent Jeanpierre and Abdel-Illah Mouaddib. *Personalized Short-Term Multi-modal Interaction for Social Robots Assisting Users in Shopping Malls.* In Adriana Tapus, Elisabeth André, Jean-Claude Martin, François Ferland and Mehdi Ammi, editors, Social Robotics, volume 9388, pages 264–274. Springer International Publishing, Cham, 2015. (Cited in page 39.)

[Itoh 2006] K. Itoh, H. Miwa, Y. Nukariya, M. Zecca, H. Takanobu, S. Roccella and et al. *Development of a Bioinstrumentation System in the Interaction between a Human and a Robot.* In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages pp. 2620–2625, Beijing, China, Oct 2006. (Cited in page 26.)

[Kahn 2008] Peter H Kahn, Nathan G Freier, Takayuki Kanda, Hiroshi Ishiguro, Jolina H Ruckert, Rachel L Severson and Shaun K Kane. *Design patterns for sociality in human-robot interaction.* In Proceedings of the 3rd ACM/IEEE international conference on Human robot interaction, pages 97–104, 2008. (Cited in page 10.)

[Kahn 2010] Peter H Kahn, Brian T Gill, Aimee L Reichert, Takayuki Kanda, Hiroshi Ishiguro and Jolina H Ruckert. *Validating interaction patterns in HRI.* In 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 183–184. IEEE, 2010. (Cited in page 10.)

[Kanda 2007] Takayuki Kanda, Rumi Sato, Naoki Saiwaki and Hiroshi Ishiguro. *A Two-Month Field Trial in an Elementary School for Long-Term Human–Robot Interaction.* IEEE Transactions on Robotics, vol. 23, no. 5, pages 962–971, 2007. (Cited in page 8.)

[Kanda 2009] Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro and Norihiro Hagita. *An affective guide robot in a shopping mall.* In ACM/IEEE international Conference on Human-Robot interaction (HRI), pages 173–180, 2009. (Cited in pages 38 and 57.)

[Kanda 2010] Takayuki Kanda, Masahiro Shiomi, Zenta Miyashita, Hiroshi Ishiguro and Norihiro Hagita. *A communication robot in a shopping mall.* IEEE Transactions on Robotics, vol. 26, no. 5, pages 897–913, 2010. (Cited in page 38.)

[Kasap 2012] Zerrin Kasap and Nadia Magnenat-Thalmann. *Building long-term relationships with virtual and robotic characters: the role of remembering.* The Visual Computer, vol. 28, no. 1, pages 87–97, 2012. (Cited in page 9.)

[Kendon 1990] Adam Kendon. Conducting interaction: Patterns of behavior in focused encounters, volume 7. CUP Archive, 1990. (Cited in pages 6, 9, and 53.)

[Keysar 1994] Boaz Keysar. *The illusory transparency of intention: Linguistic perspective taking in text.* Cognitive psychology, vol. 26, no. 2, pages 165–208, 1994. (Cited in page 83.)

[Keysar 1998] Boaz Keysar, Dale J Barr and William S Horton. *The egocentric basis of language use: Insights from a processing approach.* Current directions in psychological science, vol. 7, no. 2, pages 46–49, 1998. (Cited in page 83.)

[Keysar 2000] Boaz Keysar, Dale J Barr, Jennifer A Balin and Jason S Brauner. *Taking perspective in conversation: The role of mutual knowledge in comprehension.* Psychological Science, vol. 11, no. 1, pages 32–38, 2000. (Cited in page 82.)

[Keysar 2002] Boaz Keysar and Dale J Barr. *Self-anchoring in conversation: Why language users do not do what they'should'.* 2002. (Cited in page 83.)

[Keysar 2003] Boaz Keysar, Shuhong Lin and Dale J Barr. *Limits on theory of mind use in adults.* Cognition, vol. 89, no. 1, pages 25–41, 2003. (Cited in page 83.)

[Khambhaita 2020] Harmish Khambhaita and Rachid Alami. *Viewing Robot Navigation in Human Environment as a Cooperative Activity.* In Nancy M. Amato, Greg Hager, Shawna Thomas and Miguel Torres-Torriti, editors, Robotics Research, pages pp. 285–300. Springer International Publishing, 2020. (Cited in pages 31 and 32.)

[Kidd 2008] Cory D Kidd and Cynthia Breazeal. *Robots at home: Understanding long-term human-robot interaction.* In 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 3230–3235. IEEE, 2008. (Cited in page 9.)

[Klein 2005] Gary Klein, Paul J. Feltovich, Jeffrey M. Bradshaw and David D. Woods. Common ground and coordination in joint activity, chapter 6, pages 139–184. John Wiley & Sons, Ltd, 2005. (Cited in page 12.)

[Knapp 1973] Mark L Knapp, Roderick P Hart, Gustav W Friedrich and Gary M Shulman. *The rhetoric of goodbye: Verbal and nonverbal correlates of human leave-taking.* Communications Monographs, vol. 40, no. 3, pages 182–198, 1973. (Cited in page 7.)

[Knoblich 2011] Günther Knoblich, Stephen Butterfill and Natalie Sebanz. *Chapter three - Psychological Research on Joint Action: Theory and Data.* In Brian H. Ross, editor, Advances in Research and Theory, volume 54 of *Psychology of Learning and Motivation*, pages 59–101. Academic Press, 2011. (Cited in pages 15 and 40.)

[Kobayashi 2018] Harumi Kobayashi, Tetsuya Yasuda, Hiroshi Igarashi and Satoshi Suzuki. *Language use in joint action: The means of referring expressions.* International Journal of Social Robotics, pages 1–9, 2018. (Cited in page 12.)

[Kopp 2007] Stefan Kopp, Paul A. Tepper, Kimberley Ferriman, Kristina Striegnitz and Justine Cassell. Trading spaces: How humans and humanoids use speech and gesture to give directions, chapter 8, pages 133–160. John Wiley & Sons, Ltd, 2007. (Cited in page 38.)

[Krauss 1977] Robert M Krauss and Sam Glucksberg. *Social and nonsocial speech.* Scientific American, vol. 236, no. 2, pages 100–105, 1977. (Cited in page 82.)

[Kruse 2013] Thibault Kruse, Amit Kumar Pandey, Rachid Alami and Alexandra Kirsch. *Human-Aware Robot Navigation: A Survey.* Robotics and Autonomous Systems, vol. vol. 61, no. 12, pages pp. 1726–1743, December 2013. (Cited in page 23.)

[Kulić 2003] Dana Kulić and Elizabeth A. Croft. *Estimating Intent for Human-robot Interaction.* In IEEE International Conference on Advanced Robotics, pages pp. 810–815, 2003. (Cited in page 26.)

[Kulic 2007] D. Kulic and E. A. Croft. *Affective State Estimation for Human–Robot Interaction.* IEEE Transactions on Robotics, vol. vol. 23, no. 5, pages pp. 991–1000, 2007. (Cited in page 26.)

[Kuo 2012] I-Han Kuo. *Designing Human-Robot Interaction for Service Applications.* PhD thesis, ResearchSpace@ Auckland, 2012. (Cited in pages 10 and 11.)

[Lallement 2014] Raphaël Lallement, Lavindra De Silva and Rachid Alami. *HATP: An HTN Planner for Robotics.* In 2nd ICAPS Workshop on Planning and Robotics, Portsmouth, United States, June 2014. (Cited in pages 20 and 93.)

[Lamarre 1994] Philippe Lamarre and Yoav Shoham. *Knowledge, certainty, belief, and conditionalisation (abbreviated version).* In Principles of knowledge representation and reasoning, pages 415–424. Elsevier, 1994. (Cited in page 17.)

[Lee 2012] Min Kyung Lee, Jodi Forlizzi, Sara Kiesler, Paul Rybski, John Antanitis and Sarun Savetsila. *Personalization in HRI: A longitudinal field experiment.* In 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 319–326. IEEE, 2012. (Cited in page 9.)

[Leite 2013] Iolanda Leite, Carlos Martinho and Ana Paiva. *Social robots for long-term interaction: a survey.* International Journal of Social Robotics, vol. 5, no. 2, pages 291–308, 2013. (Cited in page 7.)

[Lemaignan 2016] S. Lemaignan, F. Garcia, A. Jacq and P. Dillenbourg. *From real-time attention assessment to "with-me-ness" in human-robot interaction.* In 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages pp. 157–164, 2016. (Cited in page 30.)

[Lemaignan 2017a] Séverin Lemaignan, Mathieu Warnier, E Akin Sisbot, Aurélie Clodic and Rachid Alami. *Artificial cognition for social human–robot interaction: An implementation.* Artificial Intelligence, vol. 247, pages 45–69, 2017. (Cited in page 89.)

[Lemaignan 2017b] Séverin Lemaignan, Matthieu Warnier, Emrah Akin Sisbot, Aurélie Clodic and Rachid Alami. *Artificial Cognition for Social Human-Robot Interaction: An Implementation.* Artificial Intelligence, vol. vol. 247, pages pp. 45–69, June 2017. (Cited in page 23.)

[Lemaignan 2018] Séverin Lemaignan, Yoan Sallami, Christopher Wallhridge, Aurélic Clodic, Tony Belpaeme and Rachid Alami. *Underworlds: cascading situation assessment for robots.* In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 7750–7757. IEEE, 2018. (Cited in pages 49 and 90.)

[Lin 2010] Shuhong Lin, Boaz Keysar and Nicholas Epley. *Reflexively mindblind: Using theory of mind to interpret behavior requires effortful attention.* Journal of Experimental Social Psychology, vol. 46, no. 3, pages 551–556, 2010. (Cited in page 83.)

[Matsumoto 2012] Takahiro Matsumoto, Satoru Satake, Takayuki Kanda, Michita Imai and Norihiro Hagita. *Do you remember that shop? computational model of spatial memory for shopping companion robots.* In Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction, pages 447–454, 2012. (Cited in pages 38, 45, and 46.)

[Mayima 2020] Amandine Mayima, Aurélie Clodic and Rachid Alami. *Toward a Robot Computing an Online Estimation of the Quality of its Interaction with its Human Partner.* In 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pages 291–298. IEEE, 2020. (Cited in page 94.)

[McNeill 2005] David McNeill. *Gesture, gaze, and ground.* In International workshop on machine learning for multimodal interaction, pages 1–14. Springer, 2005. (Cited in page 53.)

[Michael 2016] John Michael, Natalie Sebanz and Günther Knoblich. *The Sense of Commitment: A Minimal Approach.* Frontiers in Psychology, vol. vol. 6, page 1968, 2016. (Cited in page 28.)

[Michael 2017] John Michael and Alessandro Salice. *The Sense of Commitment in Human–Robot Interaction.* International Journal of Social Robotics, vol. vol. 9, no. 5, pages pp. 755–763, Nov 2017. (Cited in page 22.)

[Milliez 2014a] Grégoire Milliez, Matthieu Warnier, Aurélie Clodic and Rachid Alami. *A framework for endowing an interactive robot with reasoning capabilities about perspective-taking and belief management.* In The 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pages pp. 1103–1109, Edinburgh, United Kingdom, August 2014. (Cited in page 22.)

[Milliez 2014b] Grégoire Milliez, Matthieu Warnier, Aurélie Clodic and Rachid Alami. *A framework for endowing an interactive robot with reasoning capabilities about perspective-taking and belief management.* In 23rd international symposium on robot and human interactive communication (RO-MAN), pages 1103–1109. IEEE, 2014. (Cited in pages 81 and 91.)

[Morales 2011] Y. Morales, Satoru Satake, Takayuki Kanda and Norihiro Hagita. *Modeling environments from a route perspective.* In ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 441–448, 2011. (Cited in page 38.)

[Newell 1994] Allen Newell. Unified theories of cognition. Harvard University Press, 1994. (Cited in page 7.)

[Ogden 2001] Bernard Ogden, Kerstin Dautenhahn and Penny Stribling. *Interactional structure applied to the identification and generation of visual interactive behavior: Robots that (usually) follow the rules.* In International Gesture Workshop, pages 254–268. Springer, 2001. (Cited in pages 9 and 10.)

[Okuno 2009] Yusuke Okuno, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro and Norihiro Hagita. *Providing route directions: design of robot's utterance, gesture, and timing.* In ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 53–60. IEEE, 2009. (Cited in pages 38 and 45.)

[Olsen 2003] Dan R. Olsen and Michael A. Goodrich. *Metrics for Evaluating Human-Robot Interaction.* In PERMIS, Gaithersburg, United States, 2003. (Cited in page 30.)

[Pacherie 2008] Elisabeth Pacherie. *The phenomenology of action: A conceptual framework.* Cognition, vol. 107, no. 1, pages 179–217, 2008. (Cited in page 13.)

[Pacherie 2012] Elisabeth Pacherie. *The Phenomenology of Joint Action: Self-Agency vs. Joint-Agency.* In Axel Seemann, editor, Joint Attention: New Developments, pages 343–389. MIT Press, 2012. (Cited in pages 12, 15, and 40.)

[Panchbhai 2020] Anand Panchbhai, Tommaso Soru and Edgard Marx. *Exploring Sequence-to-Sequence Models for SPARQL Pattern Composition.* In Iberoamerican Knowledge Graphs and Semantic Web Conference, pages 158–165. Springer, 2020. (Cited in page 95.)

[Peltason 2010] Julia Peltason and Britta Wrede. *Pamini: A framework for assembling mixed-initiative human-robot interaction from generic interaction patterns.* In Proceedings of the SIGDIAL 2010 Conference, pages 229–232, 2010. (Cited in page 11.)

[Petrick 2012] Ronald PA Petrick, Mary Ellen Foster and Amy Isard. *Social state recognition and knowledge-level planning for human-robot interaction in a bartender domain.* In AAAI Workshop on Grounding Language for Physical Systems, Toronto, ON, Canada, July, 2012. (Cited in page 81.)

[Ramenzoni 2008] Verónica C Ramenzoni, Michael A Riley, Kevin Shockley and Tehran Davis. *Short article: Carrying the height of the world on your ankles: Encumbering observers reduces estimates of how high an actor can jump.* Quarterly Journal of Experimental Psychology, vol. 61, no. 10, pages 1487–1495, 2008. (Cited in page 15.)

[Rao 1995] Anand S Rao, Michael P Georgeff*et al. BDI agents: From theory to practice.* In ICMAS, volume 95, pages 312–319, 1995. (Cited in pages 16 and 17.)

[Robinson 2012] Jeffrey D. Robinson. The handbook of conversation analysis, volume 121, chapter Overall Structural Organization. John Wiley & Sons, 2012. (Cited in pages 5 and 21.)

[Rubio-Fernández 2017] Paula Rubio-Fernández. *The director task: A test of Theory-of-Mind use or selective attention?* Psychonomic bulletin & review, vol. 24, no. 4, pages 1121–1128, 2017. (Cited in page 83.)

[Rummel 1976] Rudolph J Rummel. *Understanding conflict and war: vol. 2: the conflict helix.* Bev-erly Hills: Sage, 1976. (Cited in page 5.)

[Sacks 1995] Harvey Sacks. Lectures on conversation. Wiley-Blackwell, Oxford, volumes i and ii edition edition, January 1995. (Cited in page 5.)

[Sanchez-Matilla 2020] Ricardo Sanchez-Matilla, Konstantinos Chatzilygeroudis, Apostolos Modas, Nuno Ferreira Duarte, Alessio Xompero, Pascal Frossard, Aude Billard and Andrea Cavallaro. *Benchmark for Human-to-Robot Handovers of Unseen Containers With Unknown Filling.* IEEE Robotics and Automation Letters, vol. 5, no. 2, pages 1642–1649, 2020. (Cited in page 76.)

[Sanelli 2017] Valerio Sanelli, Michael Cashmore, Daniele Magazzeni and Luca Iocchi. *Short-term human-robot interaction through conditional planning and execution.* In Proceedings of the International Conference on Automated Planning and Scheduling, volume 27, 2017. (Cited in page 7.)

[Santiesteban 2012] Idalmis Santiesteban, Sarah White, Jennifer Cook, Sam J Gilbert, Cecilia Heyes and Geoffrey Bird. *Training social cognition: from imitation to theory of mind.* Cognition, vol. 122, no. 2, pages 228–235, 2012. (Cited in page 83.)

[Sarthou 2019a] Guillaume Sarthou, Rachid Alami and Aurélie Clodic. *Semantic Spatial Representation: a unique representation of an environment based on*

an ontology for robotic applications. In Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP), 2019. (Cited in pages 48 and 51.)

[Sarthou 2019b] Guillaume Sarthou, Aurélie Clodic and Rachid Alami. *Ontologenius: A long-term semantic memory for robotic agents*. In 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pages 1–8. IEEE, 2019. (Cited in page 48.)

[Sarthou 2021] Guillaume Sarthou, Mayima Amandine, Buisan Guilhem, Belhassein Kathleen and Aurélie Clodic. *The Director Task: a Psychology-Inspired Task to Assess Cognitive and Interactive Robot Architectures*. In IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), pages 1–8. IEEE, 2021. (Cited in page 80.)

[Satake 2015a] Satoru Satake, Kotaro Hayashi, Keita Nakatani and Takayuki Kanda. *Field trial of an information-providing robot in a shopping mall*. In IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1832–1839. IEEE, 2015. (Cited in pages 45 and 48.)

[Satake 2015b] Satoru Satake, Keita Nakatani, Kotaro Hayashi, Takyuki Kanda and Michita Imai. *What should we know to develop an information robot?* PeerJ Computer Science, vol. 1, page 8, 2015. (Cited in pages 38, 45, and 80.)

[Sauppé 2014] Allison Sauppé and Bilge Mutlu. *Design patterns for exploring and prototyping human-robot interactions*. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1439–1448, 2014. (Cited in page 10.)

[Schegloff 1973] Emanuel A Schegloff and Harvey Sacks. *Opening up closings*. Semiotica, vol. 8, no. 4, pages 289–327, 1973. (Cited in pages 7 and 21.)

[Schegloff 1977] Emanuel A Schegloff, Gail Jefferson and Harvey Sacks. *The preference for self-correction in the organization of repair in conversation*. Language, vol. 53, no. 2, pages 361–382, 1977. (Cited in page 16.)

[Schegloff 1986] Emanuel A Schegloff. *The routine as achievement*. Human studies, vol. 9, no. 2-3, pages 111–151, 1986. (Cited in page 6.)

[Schegloff 2011] Emanuel A Schegloff. *Word repeats as unit ends*. Discourse Studies, vol. 13, no. 3, pages 367–380, 2011. (Cited in page 5.)

[Sebanz 2006] Natalie Sebanz, Harold Bekkering and Günther Knoblich. *Joint action: bodies and minds moving together*. Trends in cognitive sciences, vol. 10, no. 2, pages 70–76, 2006. (Cited in pages 12 and 15.)

[Sidner 2003] Candace L Sidner and Christopher Lee. *Engagement rules for human-robot collaborative interactions.* In IEEE International Conference on Systems, Man and Cybernetics., volume 4, pages 3957–3962. IEEE, 2003. (Cited in page 21.)

[Siegwart 2003] Roland Siegwart, Kai O Arras, Samir Bouabdallah, Daniel Burnier, Gilles Froidevaux, Xavier Greppin, Björn Jensen, Antoine Lorotte, Laetitia Mayor, Mathieu Meisser*et al. Robox at Expo. 02: A large-scale installation of personal robots.* Robotics and Autonomous Systems, vol. 42, no. 3-4, pages 203–222, 2003. (Cited in page 38.)

[Singamaneni 2020] Phani-Teja Singamaneni and Rachid Alami. *HATEB-2: Reactive Planning and Decision making in Human-Robot Co-navigation.* In EEE International Symposium on Robot and Human Interactive Communication (RO-MAN), pages 179–186. IEEE, 2020. (Cited in page 55.)

[Steinfeld 2006] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz and Michael Goodrich. *Common Metrics for Human-robot Interaction.* In Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-robot Interaction, pages pp. 33–40, Salt Lake City, United States, 2006. (Cited in pages 25 and 76.)

[Tabrez 2020] Aaquib Tabrez, Matthew B. Luebbers and Bradley Hayes. *A Survey of Mental Modeling Techniques in Human–Robot Teaming.* Current Robotics Reports, Aug 2020. (Cited in page 22.)

[Tanevska 2017] A. Tanevska, G. Rea F.and Sandini and A. Sciutti. *Towards an Affective Cognitive Architecture for Human-Robot Interaction for the iCub Robot.* In 1st Workshop on "Behavior, Emotion and Representation: Building Blocks of Interaction", Bielefeld, Germany, October 2017. (Cited in pages 25 and 76.)

[Tellex 2014] Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus and Nicholas Roy. *Asking for Help Using Inverse Semantics.* In Proceedings of Robotics: Science and Systems, Berkeley, USA, July 2014. (Cited in page 81.)

[Thomaz 2016] Andrea Thomaz, Guy Hoffman and Maya Çakmak. *Computational Human-Robot Interaction.* Foundations and Trends in Robotics, vol. vol. 4, no. 2-3, pages pp. 105–223, 2016. (Cited in page 23.)

[Thrun 1999] S. Thrun, M. Bennewitz, W. Burgard, A. B. Cremers, F. Dellaert, D. Fox, D. Hahnel, C. Rosenberg, N. Roy, J. Schulte and D. Schulz. *MINERVA: a second-generation museum tour-guide robot.* In Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No.99CH36288C), volume 3, pages 1999–2005, 1999. (Cited in page 38.)

[Tollefsen 2005] Deborah Tollefsen. *Let's Pretend!: Children and Joint Action.* Philosophy of the Social Sciences, vol. 35, no. 1, pages 75–97, 2005. (Cited in page 12.)

[Tomasello 2005] Michael Tomasello, Malinda Carpenter, Josep Call, Tanya Behne and Henrike Moll. *Understanding and sharing intentions: The origins of cultural cognition.* Behavioral and brain sciences, vol. 28, no. 5, pages 675–691, 2005. (Cited in page 12.)

[Triebel 2016] Rudolph Triebel, Kai Arras, Rachid Alami, Lucas Beyer, Stefan Breuers, Raja Chatila, Mohamed Chetouani, Daniel Cremers, Vanessa Evers, Michelangelo Fiore*et al. Spencer: A socially aware service robot for passenger guidance and help in busy airports.* In Field and service robotics, pages 607–622. Springer, 2016. (Cited in page 38.)

[Vesper 2010] Cordula Vesper, Stephen Butterfill, Günther Knoblich and Natalie Sebanz. *A minimal architecture for joint action.* Neural Networks, vol. 23, no. 8-9, pages 998–1003, 2010. (Cited in page 15.)

[Vesper 2011] Cordula Vesper, Robrecht PRD Van Der Wel, Günther Knoblich and Natalie Sebanz. *Making oneself predictable: Reduced temporal variability facilitates joint action coordination.* Experimental brain research, vol. 211, no. 3-4, pages 517–530, 2011. (Cited in page 15.)

[Waldhart 2019] J. Waldhart, A. Clodic and R. Alami. *Reasoning on Shared Visual Perspective to Improve Route Directions.* In 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), New Delhi, India, Oct 2019. (Cited in page 53.)

[Wooffitt 2008] Robin Wooffitt and I Hutchby. Conversation analysis. Polity, 2008. (Cited in page 16.)

[Wooldridge 1999] Michael Wooldridge. *Intelligent agents.* Multiagent systems, vol. 6, 1999. (Cited in page 44.)

[Zheng 2013] Kuanhao Zheng, Dylan F Glas, Takayuki Kanda, Hiroshi Ishiguro and Norihiro Hagita. *Designing and implementing a human–robot team for social interactions.* IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 43, no. 4, pages 843–859, 2013. (Cited in page 7.)

Abstract:

Robots will interact more and more with humans in the future and thus will need to be endowed with the pertinent abilities. We are still far from having autonomous robots among humans and able to smoothly collaborate with them but the work of this thesis is a contribution bringing the community a bit closer to this goal.

When humans collaborate to achieve a task together, numerous neurocognitive mechanisms come into play, more than we would have thought at first glance. Some of these mechanisms are also triggered in humans' minds when they interact with robots as they are essential to a successful collaboration. Therefore, it is important for roboticists designing robots that will closely interact with humans to be aware of and take into account the humans mental states and sensorimotor functions involved in controlling and smoothing collaborative task performance. However, this does not imply that robots have to be endowed with the same mechanisms since being able to collaborate with humans does not mean to imitate them. What is key to roboticists is to understand how humans work and to design robots that will adapt.

Consequently, this manuscript starts with an immersion in philosophy and social and cognitive psychology. We develop key elements for collaboration such as joint action, commitment and shared plans. Then, we explore Belief-Desire-Intention (BDI) and cognitive robotic architectures which have inspired us to design our own architecture in which, JAHRVIS — the main contribution of this thesis — endows a robot with the abilities not only to control, but also to evaluate its joint action with a human.

JAHRVIS (Joint-Action based Human-Aware supeRVISor) is what we call a supervision system, i.e., it embeds the robot high-level decisions, controls its behavior and tries to react to contingencies, always considering the human it is interacting with. It is able to do so by taking into account shared plans, human mental states, its knowledge about the current state of the environment, and human actions. The module to monitor and recognize these actions is model-based and allows to take into account a potentially unreliable perception of the actions by the robot. JAHRVIS is designed in such a way that it is generic enough to handle various kinds of tasks. It can also manage different kinds of human-robot shared plans as input: shared plans for which actions might not be allocated to an agent at planning time and objects might be referred to with a semantic query, and conditional shared plans which anticipate different possibilities for the human decision/action.

Not only JAHRVIS controls the robot contribution to a collaborative task, it also tries to evaluate if the interaction is going well or not. It is possible thanks to a set of metrics we have built and a method to aggregate them. We claim that having a robot with this ability allows it to enhance and make more pertinent its decision-making processes. The evaluation of the Quality of Interaction (QoI) relies on a model of interaction, considered at three levels: the interaction session level, the tasks level and the actions level. In future work, this granularity will allow the robot to know precisely on what level it needs to act when a low QoI is assessed.

Indeed, for instance, a task can be of poor quality while the session can still be considered as going well.

JAHRVIS has been integrated in a cognitive robotic architecture and effectively deployed to achieve several collaborative and service tasks in a real environment such as a direction-giving task in a Finnish mall and a two-agents task inspired from psychology. These tasks demonstrated the robot's abilities related to perspective-taking, planning, knowledge representation with theory of mind, manipulation, and communication.

Résumé :

Dans le futur, les robots interagiront chaque jour un peu plus avec les humains et devront donc être dotés des capacités adéquates. Nous sommes encore loin de robots autonomes parmi les humains, capables de collaborer sans problème avec eux, mais le travail de cette thèse est une contribution qui rapproche un peu plus la communauté de cet objectif.

Lorsque des personnes collaborent pour réaliser une tâche ensemble, de nombreux mécanismes neurocognitifs entrent en jeu, plus qu'il n'y paraît à première vue. Certains de ces mécanismes sont aussi activés quand un humain interagit avec un robot et non plus avec un autre humain, car ils sont essentiels à une collaboration réussie. Il est donc important que les roboticiens qui conçoivent des robots destinés à interagir étroitement avec les humains soient conscients de cela et qu'ainsi ils prennent en compte les états mentaux des humains et les fonctions sensori-motrices impliquées dans le contrôle et la fluidité de l'exécution des tâches collaboratives. Toutefois, cela ne signifie pas que les robots doivent être dotés de ces mêmes mécanismes, car être capable de collaborer avec les humains ne signifie pas les imiter. Ce qui est essentiel pour les roboticiens, c'est de comprendre comment les humains travaillent et de concevoir des robots qui s'adapteront.

Par conséquent, ce manuscrit commence par une immersion dans la philosophie et la psychologie sociale et cognitive. Nous développons les éléments clés de la collaboration tels que l'action conjointe, l'engagement et les plans partagés. Ensuite, nous explorons les architectures "croyance-désir-intention" (Belief-Desire-Intention en anglais) et les architectures robotiques cognitives qui nous ont inspirés pour concevoir notre propre architecture dans laquelle, JAHRVIS - la principale contribution de cette thèse - dote un robot des capacités de non seulement de contrôler, mais aussi d'évaluer son action conjointe avec un humain.

JAHRVIS (Joint-Action based Human-Aware supeRVISor) est ce que nous appelons un système de supervision, c'est-à-dire qu'il prend les décisions haut niveau du robot, contrôle son comportement et tente de réagir aux imprévus, en tenant toujours compte de l'humain avec lequel il interagit. Il peut le faire en se basant sur les plans partagés qu'il génère, les états mentaux de l'humain, sa connaissance de l'état actuel de l'environnement et les actions de l'humain. Le module de surveillance et de reconnaissance de ces actions est basé sur un modèle et permet au robot de prendre en compte une perception potentiellement peu fiable. JAHRVIS est conçu de manière à être suffisamment générique pour gérer différents types de tâches. Il peut également gérer différents types de plans partagés homme-robot en

entrée : des plans partagés pour lesquels les actions peuvent ne pas être attribuées à un agent au moment de la planification et pour lesquels les objets peuvent être référencés par une requête sémantique, et des plans partagés conditionnels qui anticipent différentes possibilités pour la décision/action humaine.

JAHRVIS ne se contente pas de contrôler la contribution du robot à une tâche collaborative, il essaie également d'évaluer si l'interaction se déroule bien ou non. Cela est possible grâce à un ensemble de métriques et à une méthode pour les agréger que nous avons conçu. Nous affirmons que le fait de doter un robot de cette capacité lui permet d'améliorer et de rendre plus pertinent son processus de prise de décision. L'évaluation de la qualité d'interaction (QoI) repose sur un modèle d'interaction à trois niveaux : le niveau de la session d'interaction, le niveau des tâches et le niveau des actions. Dans les travaux futurs, cette granularité permettra au robot de savoir précisément à quel niveau il doit agir lorsqu'une faible QoI est évaluée. En effet, par exemple, une tâche peut être de mauvaise qualité alors que la session peut encore être considérée comme se déroulant bien.

JAHRVIS a été intégré dans une architecture robotique cognitive et déployé efficacement pour réaliser plusieurs tâches de collaboration et de service dans un environnement réel, comme une tâche de guidage dans un centre commercial finlandais et une tâche à deux agents inspirée de la psychologie. Ces tâches ont démontré les capacités du robot en matière de prise de vue, de planification, de représentation des connaissances avec théorie de l'esprit, de manipulation et de communication.

Keywords: human-robot interaction,

Mots clés : interaction humain-robot,