Chapter 1

From Human-Human Joint Action to Human-Robot Joint Action

Contents

1.1 Joint Action Theory 1
1.1.1 Commitment 2
1.1.2 Perception and prediction 6
1.1.3 Coordination 7
1.2 How to endow a robot with Joint Action abilities 9
1.2.1 Engagement and Intention 9
1.2.2 Perspective taking and humans mental states 10
1.2.3 Coordination 11
1.3 A three levels architecture 12
1.3.1 The three levels of Pacherie 12
1.3.2 A three levels robotics architecture 13
1.3.3 Comparison to other robotics architectures 16

1.1 Joint Action Theory

A first step to endow robots with the ability to perform Joint Actions with humans is to understand how humans act together. As a working definition of Joint Action, we will use the one from [Sebanz 2006]:

> *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.*

A given number of prerequisites are needed for these individuals to achieve the so-called Joint Action. First of all, they need to agree on the change they want to bring in the environment, the conditions under which they will stay engaged in its realization and the way to do it. A number of works have studied this problematic,

relative to *commitment*, which I will develop in Sec. 1.1.1. Then, as mentioned in the definition, the individuals need to coordinate their actions in space and time. This will be studied in Sec. 1.1.3. Finally, in order to coordinate, each individual needs to be aware of the other, he needs to be able to perceive him and predict his actions. This part will be developed in Sec. 1.1.2.

1.1.1 Commitment

The first prerequisite to achieve a Joint Action is to have a *goal* to pursue and the *intention* to achieve it. Let's define what is called a *goal* and an *intention* for a single person before going to a *joint goal* and a *joint intention*.

In [Tomasello 2005], Tomasello et al. define what they call a *goal* and an *intention* and illustrate these definitions with an example and an associated figure (fig. 1.1) where a person wants to open a box.

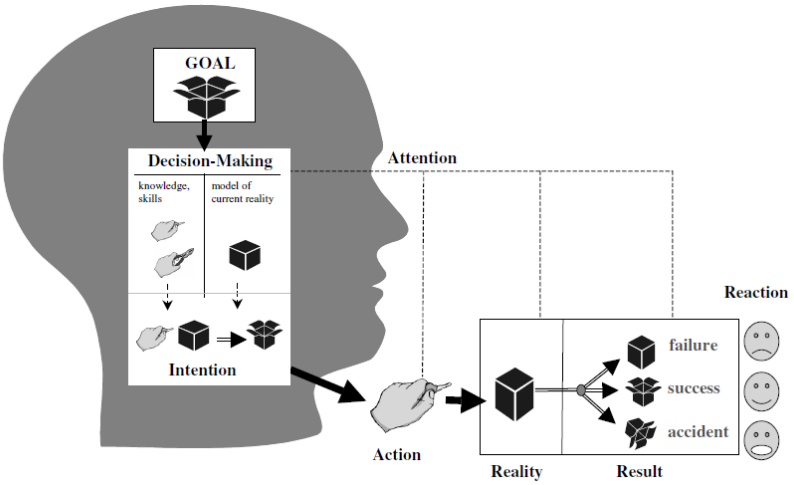

Figure 1.1: Illustrative example of an intentional action by Tomasello et al. Here the human has for *goal* for the box to be opened. He chooses a means to perform it and so forms an *intention*.

A *goal* is defined here as the representation of the desired state by the agent (in the example, the goal is an open box) and, based on Bratman's work [Bratman 1989], an *intention* is defined as an action plan the agent commits to in pursuit of a goal (in the example, the intention is to use a key to open the box). The *intention* includes both a *goal* and the means to achieve it.

In a same way, Cohen and Levesque propose in [Cohen 1991] a formal definition of what they call a *persistent goal*:

> **Definition:** An agent has a *persistent goal* relative to q to achieve p iff:
>
> 1. she believes that p is currently false;
> 2. she wants p to be true eventually;

3. it is true (and she knows it) that (2) will continue to hold until she comes to believe either that p is true, or that it will neither be true, or that q is false.

However, their definition of an *intention* differs a little from the previous one. They define an *intention* as a commitment to act in a certain mental state:

> **Definition:** An agent *intends* relative to some conditions to do an action just in case she has a persistent goal (relative to that condition) of having done the action, and, moreover, having done it, believing throughout that she is doing it.

The *intention* still includes the *goal* but here it concerns more the fact that the agent commits to achieving the goal than the way to achieve it.

Let's now apply these principles to a Joint Action. One of the best known definition of *joint intention* is the one of Bratman [Bratman 1993]:

> We intend to J if and only if:
>
> 1. (a) I intend that we J and (b) you intend that we J.
> 2. I intend that we J in accordance with and because of $1a$, $1b$, and meshing subplans of $1a$ and $1b$; you intend that we J in accordance with and because of $1a$, $1b$, and meshing subplans of $1a$ and $1b$.
> 3. 1 and 2 are common knowledge between us.

This definition is taken back and illustrated by Tomasello et al. in [Tomasello 2005] where they reuse the example of the box to open (fig 1.2).

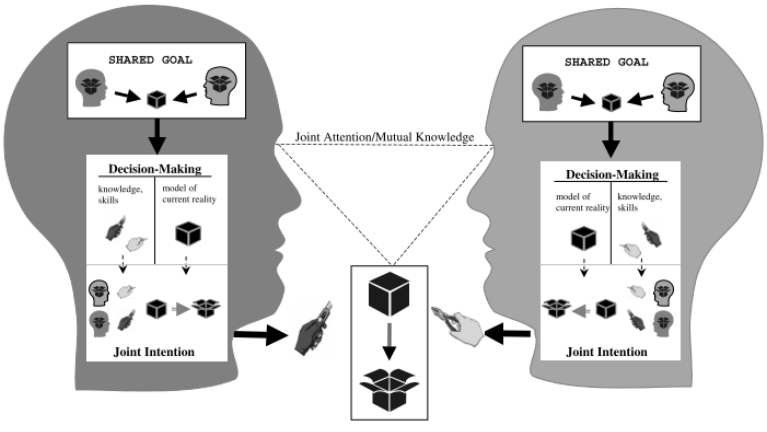

Figure 1.2: Illustrative example of a collaborative activity by Tomasello et al. Here the humans have for *shared goal* to open the box together. They choose a means to perform it which takes into account the other capabilities and so form a *joint intention*.

The *shared goal* is defined as the representation of a desired state plus the fact that it will be done in collaboration with other person(s) (in the example, they will open the box together) and a *joint intention* is defined as a collaborative plan the agents commit to in order to achieve the *shared goal* and which takes into account both agents individual plans (here an agent will hold the box with the clamp while the other open it with the cutter).

In a same way, Cohen and Levesque extend their definition of *persistent goal* and *intention* to a collaborative activity. They first define a *weak achievement goal* as:

> **Definition:** An agent has a *weak achievement goal* relative to q and with respect to a team to bring about p if either of these conditions holds:
>
> - The agent has a normal achievement goal to bring about p, that is, the agent does not yet believe that p is true and has p eventually being true as goal.
> - The agent believes that p is true, will never be true, or is irrelevant (that is, q is false), *but* has as a goal that the status of p be mutually believed by all the team members.

They then use this definition to define a *joint persistent goal*:

> **Definition:** A team of agents have a *joint persistent goal* relative to q to achieve p just in case
>
> - they mutually believe that p is currently false;
> - they mutually know they all want p to eventually be true;
> - it is true (and mutual knowledge) that until they come to mutually believe that p *is true*, that p will never be true, or that q is false, they will continue to mutually believe that they each have p as a weak achievement goal relative to q and with respect to the team.

They finally define a *joint intention* as:

> **Definition:** A team of agents *jointly intends*, relative to some escape condition, to do an action iff the members have a joint persistent goal relative of that condition of their having done the action and, moreover, having done it mutually believing throughout that they were doing it.

As previously, the definitions of Cohen and Levesque do no take into account the way to achieve the *shared goal*, however, they introduce the interesting idea that agents are also engaged to communicate about the state of the *shared goal*.

Concerning the way to achieve a *shared goal*, mentioned into the definition of the *joint intention* of Tomasello et al., Grosz and Sidner initially introduce and formalize the notion of *Shared Plan* in [Grosz 1988], which is extended in [Grosz 1999]. The key properties of their model are as follows:

1. it uses individual intentions to establish commitment of collaborators to their joint activity
2. it establishes an agent's commitments to its collaborating partners' abilities to carry out their individual actions that contribute to the joint activity
3. it accounts for helpful behavior in the context of collaborative activity
4. it covers contracting actions and distinguishes contracting from collaboration
5. the need for agents to communicate is derivative, not stipulated, and follows from the general commitment to the group activity
6. the meshing of subplans is ensured it is also derivative from more general constraints.

With their definition, each agent does not necessarily know the whole *Shared Plan* but only his own individual plan and the meshing subparts of the plan. The group has a *Shared Plan*, but no individual member necessarily has the whole *Shared Plan*.

In conclusion, the concepts concerning the commitment of agents to a collaborative activity that we will use in this thesis can be summarized as:

- A *goal* will be represented as a desired state.
- A*shared goal* will be considered as a *goal* to be achieved in collaboration with other partner(s). An agent is considered engaged in a *shared goal* if he believes the goal is currently false, he wants the goal to be true and he will not give up on the goal unless he knows that the goal is achieved, not feasible or not relevant any more and he knows that his partners are aware of it.
- A *joint intention* will include a *shared goal* and the way to realize it, represented as a *Shared Plan* which will take into account the capacities of each agent and the potential conflicts between their actions. This *Shared Plan* will not be necessarily completely known by all members of the group but all individuals will know their part of the plan and the meshing subparts.

If we apply this to the box example of Tomasello it gives us:

- "The box will be open" can be a *goal* for an agent.
- "The box will be open because we collaborate" can be a *shared goal* for several agents. Once the agents agree to achieve this goal, they will not give up until the box is open (and the other agent knows it), the box can not be opened (and the other agent knows it) or there is no more need to open the box (and the other agent knows it).

- A joint intention for two agents relative to the *shared goal* to open the box will be, for example, that the first agent go get the opener, he gives it to the second agent and then the second agent opens the box with the opener. The sequence of action <go get the opener, give it, open the box> is the Shared Plan. The second agent does not need details concerning the part "go get the opener" while the first agent does not need details concerning the part "open the box".

1.1.2 Perception and prediction

One important thing for an agent when performing a Joint Action is to be able to perceive and predict the actions of his partner and their effects. Based on the works in [Sebanz 2006], [Pacherie 2011] and [Obhi 2011] we identified several necessary abilities for this predictions:

Joint attention: The capacity for an agent to share focus with his partner allows to share a representation of objects and events. It brings a better understanding of the other agent's knowledge and where his attention is focused and so, it helps the prediction of his possible next actions. Moreover, there should be a mutual manifestation of this joint attention, meaning that we should show that we share the other attention.

Action observation: Several studies have shown that when someone observes another person executing an action, a corresponding representation of the action is formed for the observer [Rizzolatti 2004]. This is done by what has been called the *mirror-neuron* system. This behavior allows the observer to predict the outcomes of the actor's action.

Co-representation: An agent needs to have a representation of his partner, including his goal, his capacities and the social rules he is following. This representation also includes the knowledge of the partner on the Shared Plan, especially on the actions attributed to him. Having this representation will help to predict his future actions. For example, as a pedestrian knows that the car drivers follow the traffic regulations, he will be able to predict that they will stop if he sees a red traffic light.

Agency: Sometimes, when there is a close link between an action performed by oneself and an action performed by someone else, it can be hard to distinguish who caused a particular effect. The capacity to attribute the effects to the right actor is called the sense of *Agency*. This sense of *Agency* is important in Joint Action in order to correctly predict the effects of each action.

Based on the same works as before and on [Sebanz 2009], we can list several kinds of predictions to support Joint Action which can be done thanks to the abilities

described previously :

- **What:** A first one is to predict what an agent will do. Two kinds of predictions, described in [Pacherie 2011], can be distinguished here:
  - *action-to-goal:* this is supported by the *mirror-neuron* system introduced before. Here the word goal designates the goal of an action, its purpose. The idea is that by observing an action, it is possible to predict its goal. For example, if we observe someone extending his arm toward an object we can predict that he will pick the object.
  - *goal-to-action:* here the word goal designates the goal of a task, as defined in the previous subsection. Knowing this goal, it can be easy to predict which action an agent will perform.
- **When:** another prediction which is necessary is the timing of an action. Knowing when an action will occur and how long it will take allows for a better coordination in time.
- **Where:** a Joint Action usually takes place in a shared space. It is therefore necessary to predict the future position of the partner and his actions in order to coordinate in space.

1.1.3 Coordination

The predictions discussed previously allow agents to coordinate during Joint Action. Two kinds of coordination are defined in [Knoblich 2011] that both support Joint Action.

Emergent coordination: It is a coordinated behavior which occurs unintentionally, independently of any joint plan or common knowledge and due to perception-action couplings. Four types of sources of emergent coordination can be distinguished:

- *Entrainment:* Entrainment is a process that leads to temporal coordination of two actors' behavior, in particular, synchronization, even in the absence of a direct mechanical coupling. It is the case, for example, for two people seating in rocking chairs involuntary synchronizing their rocking frequencies [Richardson 2007].
- *Affordances:* An object affordance represents the opportunities that an object provides to an agent for a certain action repertoire [Gibson 1977]. For example, the different ways to grab a mug. Two kinds of affordances can lead to an emergent coordination: *common affordances* and *joint affordances*. When several agents have the same action repertoire and perceive the same object they have a *common affordance*. This *common affordance* can lead the agents to execute the same action. When an object has affordances for two or more

peoples collectively, the agents have to synchronize for an action to occur. This is what is called *joint affordances*. For example, a long two-handled saw affords cutting for two people acting together but not for either of them acting individually.

- *Perception-action matching:* As discussed before, observing an action activates corresponding representation in the observer's mind. This process can lead to involuntary mimicry of the observed action. Consequently, if two persons observe the same action, they can have the same reaction to mimic the action.

- *Action simulation:* The internal mechanisms activated during action observation not only allow to mimic the action but also to predict the effects of this action. If two people observe the same action and so predict the same effects, they can consequently have the same reaction. For example, two persons seeing the same object falling will have the same reaction to try to catch it.

Planned coordination: While emergent coordination is unintentional, planned coordination requires for agents to plan their own actions in relation to Joint Action and others' actions.

One way for an agent to intentionally coordinate during Joint Action is to change his behavior compared to when he is acting alone. These changes of behavior are called *coordination smoothers* in [Vesper 2010] and can be of several types:

- Making our behavior more predictable by doing for example wider or less variable movements

- Structuring our own task in order to reduce the need of coordination. For example sharing the space or working turn by turn.

- Producing coordination signals like looking someone who should act or counting down.

- Changing the way we use an object by using an affordance more appropriate to a shared use.

Another way to coordinate is through communication. Indeed, Clark argues that two or more persons cannot perform a Joint Action without communicating [Clark 1996]. Here the word communication includes both verbal and non-verbal communication. Clark also defines what he calls the *common ground*: when two agents communicate, they necessarily have common knowledge and conventions. Moreover, when communicating, it is important to not only send a message but also to assure that the message has been understood as the sender intends it to be. This process to make the sender and the receiver mutually believe that the message has been understood well enough for current purposes is called *grounding*.

In conclusion, in order to smoothly perform a Joint Action, an agent needs to:

- Develop sufficient perception and prediction abilities in order to coordinate in space and time. It needs to be done from basic motor commands to high level decisions.
- Produce coordination signals understandable by his partners in order for them to predict the agent behavior.
- Unsure that the signals he sends are well received by his partners.

1.2 How to endow a robot with Joint Action abilities

In this part we will do an overview of how the theory on human-human Joint Action can be applied to human-robot Joint Action. Following to what has been discussed on commitment, we will first see in Sec. 1.2.1 how the robot can engage in Joint Action and understand the intention of its human partners. Then, we will see in Sec. 1.2.2 how the robot perceives the humans and can predict their actions by taking into account their perspectives and mental states. We will also see how the robot can coordinate during Joint Action in Sec. 1.2.3.

The parts which are linked to the work presented in this thesis will be more developed in the corresponding chapters.

1.2.1 Engagement and Intention

As for humans, robots need to be able to engage in Joint Action. A first prerequisite is to choose a goal to perform. This goal can be imposed by a direct order of the user, however, the robot also needs to be able to pro actively propose its help whenever a human needs it. To do so, the robot needs to be able to infer high-level goals by observing and reasoning on its human partners' activities. This process is called plan recognition or, when a bigger focus is put on human-robot interaction aspects, intention recognition. Many works have been done concerning plan recognition using approaches such as classical planning [Ramırez 2009], probabilistic planning [Bui 2003] or logic-based techniques [Singla 2011]. Concerning intention recognition, works such as [Breazeal 2009] and [Baker 2014] take into account theory of mind aspects to deduce what the human is doing.

When direct orders have been received and humans intentions recognized, the robot needs to choose which goal to perform, also taking into account its own resources. This problem has not been addressed as a whole in the literature, however, some similar works can be seen as partial answers. For example, some deliberation systems allow to solve problems with multiple goals taking into account resources such as time [Georgeff 1987, Ghallab 1994, Lemai 2004] or energy level [Rabideau 1999].

Once the robot is engaged in a Joint Action, it needs to be able to monitor other agents engagement. Indeed, it needs to understand if, for a reason, a human aborts the current goal and reacts accordingly. This can be done using gaze cues and

gestures [Rich 2010], postures [Sanghvi 2011] but also context and humans mental states [Salam 2015].

Finally, once a goal is chosen, the robot needs to be able to establish a Shared Plan to achieve it with its human partners. Several works have been done in task planning to take into account the human [Cirillo 2010, Lallement 2014]. They allow the robot to reduce resource conflicts [Chakraborti 2016], take divergent beliefs into account [Guitton 2012, Talamadupula 2014] or promote stigmergic collaboration for agents in co-habitation [Chakraborti 2015]. Once the plan computed, the robot needs to be able to share/negotiate it with its partners. Several studies have been reported on how to communicate these plans. Some researchers studied how a system could acquire knowledge on plan decomposition from a user [Mohseni-Kabir 2015] and how dialog can be used to teach new collaborative plans to the robot and to modify these plans [Petit 2013]. In [Sorce 2015], the system is able to learn a plan from a user and transmit it to another user and in [Allen 2002] a computer agent is able to construct a plan in collaboration with a user. Finally, [Milliez 2016] presents a system where the robot shares the plan with a level of details which depends on the expertise of the user.

1.2.2 Perspective taking and humans mental states

One of the first difference between a human and a robot is the way they perceive the world. To perceive its environment, the robot uses sensors to recognize and localize entities. These sensors return positions and orientations in the form of coordinates (x, y, z, θ, ϕ, ψ). On the other hand, humans use relations between objects to describe their positions (e.g. the mug is on the kitchen table, oriented toward the window). To understand the human references and to generate understandable utterances, the robot needs therefore to build a semantic representation of the world, based on the geometric data it collects from sensors. This process is called *grounding* and has been developed in several works as [Mavridis 2005] or [Lemaignan 2012].

However having its own semantic representation of the world is not enough for the robot, it also needs to take into account the point of view of its partners in order to better understand their goals and actions. To do so, the robot does what is called *perspective taking*, it constructs a representation of the world from the humans perspectives [Breazeal 2006, Milliez 2014]. This ability can be used by the robot to choose its actions in order to influence others mental states [Gray 2014], solve ambiguous situations [Ros 2010] or to better interact during dialogue [Ferreira 2015].

One important application of *perspective taking* is human action recognition. Indeed, knowing what others are aware of is a first step to understand what they are doing. Then, the action recognition can be done based on Partially Observed Markov Decision Processes (POMDP) and Dynamic Bayesian Networks (DBN) [Baker 2014] or inverse reinforcement learning [Nagai 2015].

This subject will be more developed and we will see, in Chapter **??**, how we use *perspective taking* to estimate mental states of the humans concerning the Shared Plan in order to improve their execution.

1.2.3 Coordination

One of the most important and difficult challenges during human robot Joint Action is to coordinate. The problem appears at different levels during Joint Action.

At a higher level, the humans and the robot need to coordinate their actions to fluidly execute the Shared Plan. The robot needs to execute its actions at the right time, monitor the ones of its partners and correctly communicate when needed. Several systems have been developed to do so, has *Chaski* [Shah 2011], a task-level executor which uses insights from human-human teaming in order to minimize human idle time or *Pike* an online executive that unifies intention recognition and plan adaptation to deal with temporal uncertainties during Shared Plan execution [Karpas 2015]. A part of the work presented in the thesis is the extension of *SHARY*, a supervisor allowing to execute Shared Plans into a complete human-aware architecture [Clodic 2009]. We will notably see in Chapter **??** how we extend it in order to execute flexible Shared Plans where part of the decisions are let to the execution.

One of the key aspects at this level of coordination is verbal and non-verbal communication. Concerning verbal communication, there are two ways to consider dialogue. The first one consists on seeing dialogue as a Joint Action. The second one is to see it as a tool for Joint Action. In practice, dialogue can be both, and, as developed in [Clark 1996], there can be Joint Actions in Joint Actions. Several works in robotics developed modules allowing the robot to dialogue with humans in support to Joint Action [Roy 2000, Lucignano 2013, Ferreira 2015]. To support dialogue and Joint Action in general, non verbal communication is very important. Its benefit has been shown for human-robot interaction [Breazeal 2005] and ways to perform it have been studied, principally concerning gaze cues [Boucher 2010, Mutlu 2009] but also postures [Hart 2014]. However, there are few works which study the use of non-verbal behavior during human-robot Joint Action where both partners are acting. This subject and the associated literature will be more developed in Chapter **??**.

At a lower level, the robot needs to coordinate with its partners during action execution. To execute a task, the robot can be led to perform actions in collaboration with one or several humans. The principal action studied in HRI is handover, an action which seems simple as we do it in every day life but which, in fact, raises a number of challenges as, among others, approaching the other person [Walters 2007], finding an acceptable posture to give the object [Cakmak 2011, Mainprice 2012] or releasing the object with the good timing [Mason 2005]. But, the robot also needs to coordinate when it is executing an action on its own. Indeed, it needs to share space and resources and its actions need to be understandable enough for its partners. To do so, the robot not only has to execute its actions in an efficient way, but also in a legible, acceptable and predictable way. This process can be compared to the *coordination smoothers* described in [Vesper 2010] and one way to do it is through human-aware motion planing [Sisbot 2012, Kruse 2013].

1.3 A three levels architecture

We saw previously the different prerequisites to Joint Action, both between humans and in HRI. We will see now how the monitoring of Joint Action is organized around three different levels first with the theory of Pacherie concerning humans Joint Actions in Sec. 1.3.1 and then in the LAAS robotics architecture in Sec. 1.3.2.

1.3.1 The three levels of Pacherie

As for the prerequisite of Joint Action, we will first introduce the concepts developed by Pacherie on Action and then, extend them to Joint Action. Pacherie argues in [Pacherie 2008] that intention in Action is composed of three levels which all have a specific role to play and which are organized as in fig. 1.3.

Distal Intention: This is the highest level of intention. In a first time, this level is in charge of forming an intention to act. It means that it is in charge of choosing a goal, a time to execute it and finding a plan to achieve it. Then, once the time comes to execute the plan, this level has to ensure its good execution. To do so, Pacherie takes back the definition of what is called *rational guidance and control* [Buekens 2001]. This control takes two forms: 'tracking control' where we ensure that each successive step in the action plan is successfully implemented before moving to the next step and 'collateral control' where we control for the side effects of accomplishing an action.

Proximal Intention: This level inherits an action plan from the *Distal Intention*. Its responsibility is, first, to anchor the received action plan which is defined in an abstract way in the situation of the action. It needs to integrate conceptual information about the intended action inherited from the *Distal Intention* with perceptual information about the current situation to yield a more definite representation of the action to be performed. Then, this level has to ensure that the imagined actions become current through situational control of their unfolding.

Motor Intention: This is the lowest level of intention. As for the other levels, it first has to make choices and then to monitor their executions. At this level, these choices concern motor commands, which are the physical ways to achieve the action inherited from the *Proximal Intention*.

In [Pacherie 2011], Pacherie extends these three levels to Joint Action. In the same way as before, these three new levels coexist at the same time, each one controlling the Joint Action at a different level.

Shared Distal Intention: Where *Distal Intention* was responsible for intention, *Shared Distal Intention* is responsible for joint intention. When performing a Joint Action, this level is the one responsible for the shared goal and the Shared Plan.

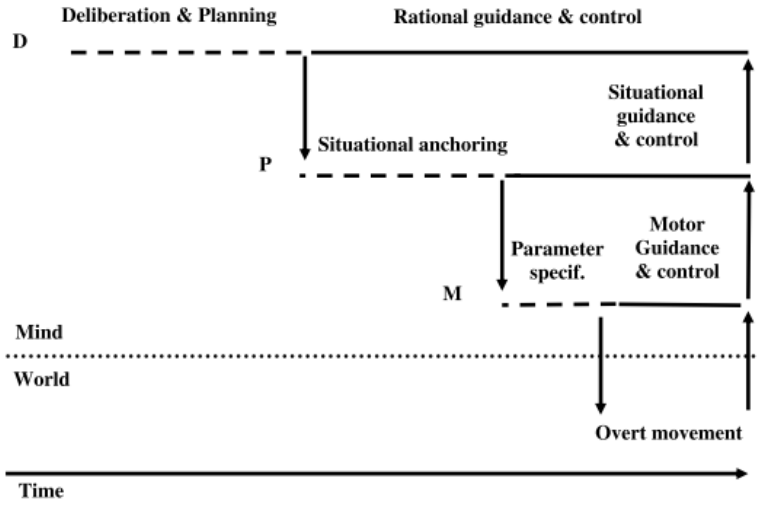

Figure 1.3: The intentional cascade by Pacherie. Distal, Proximal and Motor intentions coexist at the same time, each one controlling the action at a different level.

As said in Sec. 1.1.1, the agent does not have a whole representation of the Shared Plan here and part of his representation will be executed by someone else.

Shared Proximal Intention: This level has the same responsibilities as *Proximal Intention*, however, the anchoring of the action plan needs to take care of the Joint Action partners and to be done in a coordinated way. During the monitoring part, the choices made previously need to be adapted to the others' behavior.

Coupled Motor Intention: As for *Motor Intention*, this level is responsible for the motor commands of the agent. During Joint Action, this level will be the one responsible for precise spatio-temporal coordination for the actions which need it (e.g. holding an object together).

1.3.2 A three levels robotics architecture

Ten years before Pacherie came with her action theory with three levels, the field of autonomous robotics was trying to build architectures and was already intuitively designing three similar levels. A first implemented architecture for autonomous robots is presented in [Alami 1998], organized around these three levels (fig. 1.4).

Decision level: This level can be compared to the *Distal Intention* level of Pacherie. It is the one responsible for producing a task plan and supervising it. It sends actions to execute and receives reports from the *execution level*.

Execution level: This level can be compared to the *Proximal Intention* level of Pacherie. It receives from the *decision level* the sequence of actions to be executed

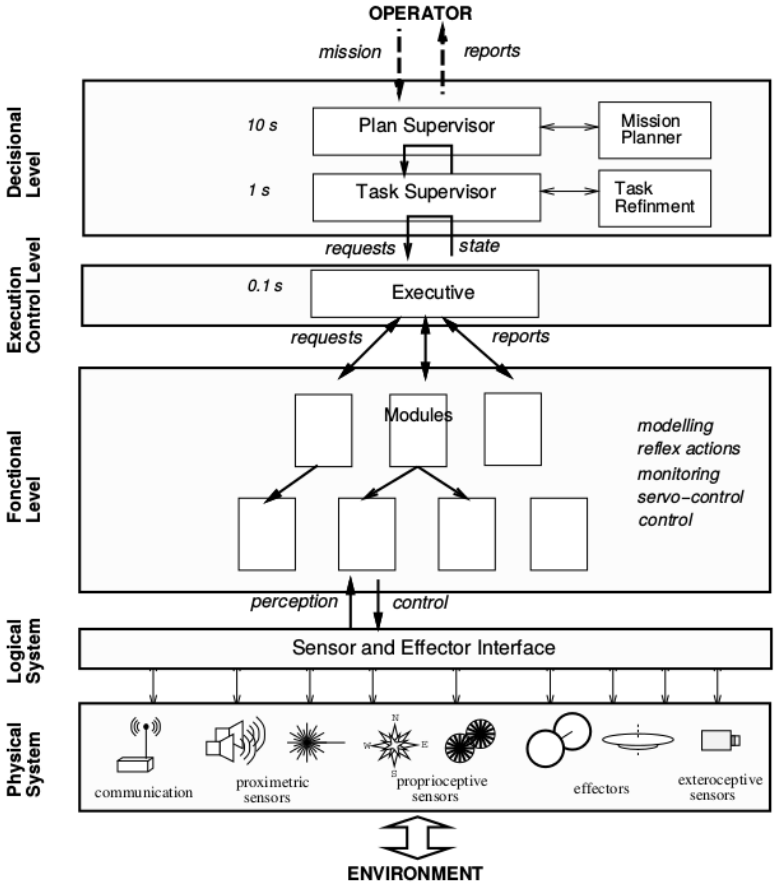

Figure 1.4: One of the first architectures for autonomous robots. The architecture is divided in three main parts: decision, execution and functional levels.

and selects, parametrizes and synchronizes dynamically the adequate functions of the *functional level*.

Functional level: This level can be compared to the *Motor Intention* level of Pacherie. It includes all the basic robot action and perception capacities (motion planning, vision, localization, tracking motion control...).

In the past years, this architecture has been developed and adapted to the field of HRI. In recent works, we presented in [Devin 2016] a theoretical version of the architecture adapted to human-robot Joint Action and still based on the three levels of Pacherie (fig. 1.5). The implemented version of this architecture will be presented in Chapter **??**, where my contribution in the architecture will also be highlighted.

Distal level: As for *Shared Distal Intention*, this level is responsible for goals and Shared Plans management. At this level, the robot is supposed to reason on its environment with high level representations. To do so, the robot is equipped

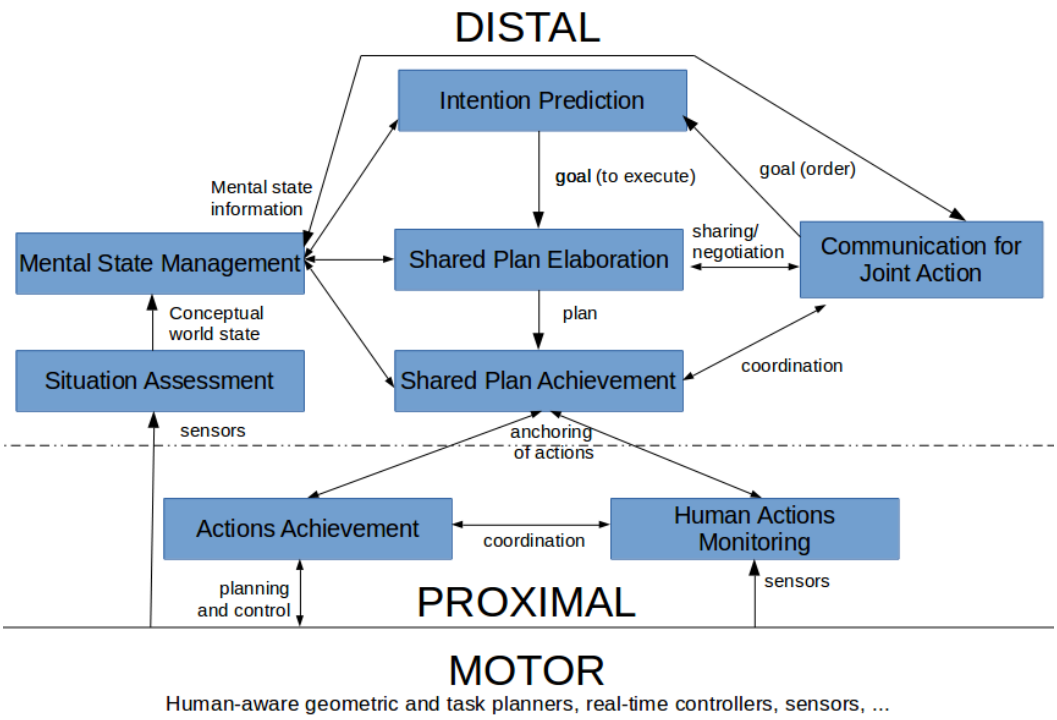

Figure 1.5: Recent architecture for human-robot Joint Action. The architecture is organized in three levels corresponding to the ones defined by Pacherie.

with a **Situation Assessment** module which builds a symbolic representation of the robot environment. To be able to also reason about the humans knowledge, the robot is equipped with a **Mental State Management** module which constantly estimates humans mental states. With this information, the **Intention prediction** module is able to estimate humans intention and if the robot should propose its help or not. This module determines the goal of the robot and allows, during its execution, to monitor other agents engagement. Once the goal chosen, the **Shared Plan Elaboration** module allows the robot to construct and negotiate a Shared Plan to achieve the goal. Then, the **Shared Plan Achievement** module monitors the good execution of this Shared Plan. The last part of this level is the **Communication for Joint Action** module which allows the robot to verbally and non-verbally communicate during Joint Action.

Proximal level: As for *Shared Proximal Intention*, this level is in charge of anchoring the Shared Plan actions in the current situation. This level is composed of two parts: the **Actions Achievement** module which allows to call the adequate motor modules at the right time in order to perform robot actions and the **Human Actions Monitoring** module which allows to recognize and interpret humans actions with regard to the Shared Plan. These two modules communicate in order to coordinate robot actions to the humans ones.

Motor level: As for *Coupled Motor Intention*, this level is in charge of motor commands of the robot. This level includes all modules allowing to control the robot actuators and interprets data from sensors. These modules, or at least a part of them, also take into account the humans as, for example, the human-aware geometric task and motion planner.

1.3.3 Comparison to other robotics architectures

We saw previously that human-robot interaction is a very complex field with many interesting subjects to study. As a consequence, few architectures allow the robot to execute tasks with humans in a fully autonomous way.

In [Baxter 2013], Baxter et al. present a cognitive architecture built around DAIM (The Distributed Associative Interactive Memory), a memory component which allows the robot to classify the humans behavior. This architecture allows the robot to fluently align its behavior with the human while memorizing data on the interaction. However, several key aspects of human-robot Joint Action as human-aware action execution or theory of mind are missing in this architecture.

Another cognitive architecture is presented in [Trafton 2013]. ACT-R/E, a cognitive architecture based on the ACT-R architecture, is used for human-robot interaction tasks. The architecture aims at simulating how humans think, perceive and act in the world. ACT-R/E has been tested in different scenarios, such as theory of mind and hide and seek, to show its capacity of modeling human behaviors and tought. This architecture has a big focus on the theory of mind and decisional aspects letting less space to the human-aware action execution or understanding, which are also big HRI challenges.

[Beetz 2010] proposes a cognitive architecture called CRAM (Cognitive Robot Abstract Machine) that integrates KnowRob [Tenorth 2013], a knowledge processing framework based on Prolog. CRAM is a very complete architecture dealing with problems such as manipulation, perception, plans or beliefs management. However, this architecture is more designed for a robot acting alone than a robot acting in collaboration with a human. Consequently, the architecture misses some key Joint Action aspects such as communication or humans actions monitoring.

An architecture based on inverse and forward models is presented in [Demiris 2006]. This architecture integrates interesting mechanisms which use human point of view to achieve learning. However, these aspects are limited to action recognition and execution and the architecture does not allow to deal with higher level decisional issues.

All of these architectures are really interesting and sharp in their respective predilection domains. However, even if our architecture may lack of some aspects as learning or memory management, its aim is to regroup a major part of the human-robot Joint Action aspects from higher level (intention management and decisional process) to lower level (human-aware execution, coordination and perception). Moreover, the architecture has been conceived in a modular enough way to allow the addition of new modules.

Bibliography

[Alami 1998] Rachid Alami, Raja Chatila, Sara Fleury, Malik Ghallab and Félix Ingrand. *An architecture for autonomy.* The International Journal of Robotics Research, vol. 17, no. 4, pages 315–337, 1998. (Cited in page 13.)

[Allen 2002] James Allen and George Ferguson. *Human-machine collaborative planning.* In Third International NASA Workshop on Planning and Scheduling for Space, 2002. (Cited in page 10.)

[Baker 2014] Chris L Baker and Joshua B Tenenbaum. *Modeling human plan recognition using bayesian theory of mind.* Plan, activity, and intent recognition: Theory and practice, pages 177–204, 2014. (Cited in pages 9 and 10.)

[Baxter 2013] Paul E Baxter, Joachim de Greeff and Tony Belpaeme. *Cognitive architecture for human–robot interaction: towards behavioural alignment.* Biologically Inspired Cognitive Architectures, vol. 6, pages 30–39, 2013. (Cited in page 16.)

[Beetz 2010] Michael Beetz, Lorenz Mösenlechner and Moritz Tenorth. *CRAM—A Cognitive Robot Abstract Machine for everyday manipulation in human environments.* In Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on, pages 1012–1017. IEEE, 2010. (Cited in page 16.)

[Boucher 2010] Jean-David Boucher, Jocelyne Ventre-Dominey, Peter Ford Dominey, Sacha Fagel and Gerard Bailly. *Facilitative effects of communicative gaze and speech in human-robot cooperation.* In Proceedings of the 3rd international workshop on Affective interaction in natural environments, pages 71–74. ACM, 2010. (Cited in page 11.)

[Bratman 1989] Michael E Bratman. *Intention and personal policies.* Philosophical perspectives, vol. 3, pages 443–469, 1989. (Cited in page 2.)

[Bratman 1993] Michael E Bratman. *Shared intention.* Ethics, vol. 104, no. 1, pages 97–113, 1993. (Cited in page 3.)

[Breazeal 2005] Cynthia Breazeal, Cory D Kidd, Andrea Lockerd Thomaz, Guy Hoffman and Matt Berlin. *Effects of nonverbal communication on efficiency and robustness in human-robot teamwork.* In Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on, pages 708–713. IEEE, 2005. (Cited in page 11.)

[Breazeal 2006] Cynthia Breazeal, Matt Berlin, Andrew Brooks, Jesse Gray and Andrea L Thomaz. *Using perspective taking to learn from ambiguous demonstrations.* Robotics and autonomous systems, vol. 54, no. 5, pages 385–393, 2006. (Cited in page 10.)

[Breazeal 2009] Cynthia Breazeal, Jesse Gray and Matt Berlin. *An embodied cognition approach to mindreading skills for socially intelligent robots.* The International Journal of Robotics Research, vol. 28, no. 5, pages 656–680, 2009. (Cited in page 9.)

[Buekens 2001] FAI Buekens, X Vanmechelen and K Maessen. *Indexicaliteit en dynamische intenties.* 2001. (Cited in page 12.)

[Bui 2003] Hung Hai Bui. *A general model for online probabilistic plan recognition.* In IJCAI, volume 3, pages 1309–1315, 2003. (Cited in page 9.)

[Cakmak 2011] Maya Cakmak, Siddhartha S Srinivasa, Min Kyung Lee, Jodi Forlizzi and Sara Kiesler. *Human preferences for robot-human hand-over configurations.* In Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on, pages 1986–1993. IEEE, 2011. (Cited in page 11.)

[Chakraborti 2015] Tathagata Chakraborti, Gordon Briggs, Kartik Talamadupula, Matthias Scheutz, David Smith and Subbarao Kambhampati. *Planning for serendipity-altruism in human-robot cohabitation.* In ICAPS Workshop on Planning and Robotics (PlanRob), 2015. (Cited in page 10.)

[Chakraborti 2016] Tathagata Chakraborti, Yu Zhang, David E Smith and Subbarao Kambhampati. *Planning with resource conflicts in human-robot cohabitation.* In AAMAS, 2016. (Cited in page 10.)

[Cirillo 2010] Marcello Cirillo, Lars Karlsson and Alessandro Saffiotti. *Human-aware task planning: an application to mobile robots.* IST, 2010. (Cited in page 10.)

[Clark 1996] Herbert H Clark. *Using language. 1996.* Cambridge University Press: Cambridge, vol. 952, pages 274–296, 1996. (Cited in pages 8 and 11.)

[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 11.)

[Cohen 1991] Philip R Cohen and Hector J Levesque. *Teamwork.* Nous, vol. 25, no. 4, pages 487–512, 1991. (Cited in page 2.)

[Demiris 2006] Yiannis Demiris and Bassam Khadhouri. *Hierarchical attentive multiple models for execution and recognition of actions.* Robotics and autonomous systems, vol. 54, no. 5, pages 361–369, 2006. (Cited in page 16.)

[Devin 2016] Sandra Devin, Grégoire Milliez, Michelangelo Fiore, Aurélie Clodic and Rachid Alami. *Some essential skills and their combination in an architecture for a cognitive and interactive robot.* arXiv preprint arXiv:1603.00583, 2016. (Cited in page 14.)

[Ferreira 2015] Emmanuel Ferreira, Grégoire Milliez, Fabrice Lefèvre and Rachid Alami. *Users' belief awareness in reinforcement learning-based situated human–robot dialogue management.* In Natural Language Dialog Systems and Intelligent Assistants, pages 73–86. Springer, 2015. (Cited in pages 10 and 11.)

[Georgeff 1987] Michael P Georgeff and Amy L Lansky. *Reactive reasoning and planning.* In AAAI, volume 87, pages 677–682, 1987. (Cited in page 9.)

[Ghallab 1994] Malik Ghallab and Hervé Laruelle. *Representation and Control in IxTeT, a Temporal Planner.* In AIPS, volume 1994, pages 61–67, 1994. (Cited in page 9.)

[Gibson 1977] James J Gibson. *Perceiving, acting, and knowing: Toward an ecological psychology.* The Theory of Affordances, pages 67–82, 1977. (Cited in page 7.)

[Gray 2014] Jesse Gray and Cynthia Breazeal. *Manipulating mental states through physical action.* International Journal of Social Robotics, vol. 6, no. 3, pages 315–327, 2014. (Cited in page 10.)

[Grosz 1988] Barbara J Grosz and Candace L Sidner. *Plans for discourse.* Technical Report, DTIC Document, 1988. (Cited in page 4.)

[Grosz 1999] Barbara J Grosz and Sarit Kraus. *The evolution of SharedPlans.* In Foundations of rational agency, pages 227–262. Springer, 1999. (Cited in page 4.)

[Guitton 2012] Julien Guitton, Matthieu Warnier and Rachid Alami. *Belief Management for HRI Planning.* European Conf. on Artificial Intelligence, 2012. (Cited in page 10.)

[Hart 2014] Justin W Hart, Brian Gleeson, Matthew Pan, AJung Moon, Karon MacLean and Elizabeth Croft. *Gesture, gaze, touch, and hesitation: Timing cues for collaborative work.* In HRI Workshop on Timing in Human-Robot Interaction, Bielefeld, Germany, 2014. (Cited in page 11.)

[Karpas 2015] Erez Karpas, Steven James Levine, Peng Yu and Brian C Williams. *Robust Execution of Plans for Human-Robot Teams.* In ICAPS, pages 342–346, 2015. (Cited in page 11.)

[Knoblich 2011] Günther Knoblich, Stephen Butterfill and Natalie Sebanz. *3 Psychological research on joint action: theory and data.* Psychology of Learning and Motivation-Advances in Research and Theory, vol. 54, page 59, 2011. (Cited in page 7.)

[Kruse 2013] Thibault Kruse, Amit Kumar Pandey, Rachid Alami and Alexandra Kirsch. *Human-aware robot navigation: A survey.* Robotics and Autonomous Systems, vol. 61, no. 12, pages 1726–1743, 2013. (Cited in page 11.)

[Lallement 2014] Raphaël Lallement, Lavindra de Silva and Rachid Alami. *HATP: An HTN Planner for Robotics.* CoRR, 2014. (Cited in page 10.)

[Lemai 2004] Solange Lemai and Félix Ingrand. *Interleaving temporal planning and execution in robotics domains.* In AAAI, volume 4, pages 617–622, 2004. (Cited in page 9.)

[Lemaignan 2012] Séverin Lemaignan, Raquel Ros, E Akin Sisbot, Rachid Alami and Michael Beetz. *Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction.* International Journal of Social Robotics, vol. 4, no. 2, pages 181–199, 2012. (Cited in page 10.)

[Lucignano 2013] Lorenzo Lucignano, Francesco Cutugno, Silvia Rossi and Alberto Finzi. *A dialogue system for multimodal human-robot interaction.* In Proceedings of the 15th ACM on International conference on multimodal interaction, pages 197–204. ACM, 2013. (Cited in page 11.)

[Mainprice 2012] Jim Mainprice, Mamoun Gharbi, Thierry Siméon and Rachid Alami. *Sharing effort in planning human-robot handover tasks.* In RO-MAN, 2012 IEEE, pages 764–770. IEEE, 2012. (Cited in page 11.)

[Mason 2005] Andrea H Mason and Christine L MacKenzie. *Grip forces when passing an object to a partner.* Experimental brain research, vol. 163, no. 2, pages 173–187, 2005. (Cited in page 11.)

[Mavridis 2005] Nikolaos Mavridis and Deb Roy. *Grounded situation models for robots: Bridging language, perception, and action.* In AAAI-05 workshop on modular construction of human-like intelligence, 2005. (Cited in page 10.)

[Milliez 2014] 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 Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on, pages 1103–1109. IEEE, 2014. (Cited in page 10.)

[Milliez 2016] Grégoire Milliez, Raphaël Lallement, Michelangelo Fiore and Rachid Alami. *Using human knowledge awareness to adapt collaborative plan generation, explanation and monitoring.* In HRI ACM/IEEE, 2016. (Cited in page 10.)

[Mohseni-Kabir 2015] A. Mohseni-Kabir, C. Rich, S. Chernova, C. L. Sidner and D. Miller. *Interactive Hierarchical Task Learning from a Single Demonstration.* In HRI, ACM/IEEE, 2015. (Cited in page 10.)

[Mutlu 2009] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro and Norihiro Hagita. *Footing in human-robot conversations: how robots might shape participant roles using gaze cues.* In Proceedings of the 4th ACM/IEEE international conference on Human robot interaction, pages 61–68. ACM, 2009. (Cited in page 11.)

[Nagai 2015] Takayuki Nagai, Kasumi Abe, Tomoaki Nakamura, Natsuki Oka and Takashi Omori. *Probabilistic modeling of mental models of others.* In Robot and Human Interactive Communication (RO-MAN), 2015 24th IEEE International Symposium on, pages 89–94. IEEE, 2015. (Cited in page 10.)

[Obhi 2011] Sukhvinder S Obhi and Natalie Sebanz. *Moving together: toward understanding the mechanisms of joint action*, 2011. (Cited in page 6.)

[Pacherie 2008] Elisabeth Pacherie. *The phenomenology of action: A conceptual framework.* Cognition, vol. 107, no. 1, pages 179–217, 2008. (Cited in page 12.)

[Pacherie 2011] Elisabeth Pacherie. *The Phenomenology of Joint Action: Self-Agency versus Joint Agency.* Joint attention: New developments in psychology, philosophy of mind, and social neuroscience, page 343, 2011. (Cited in pages 6, 7, and 12.)

[Petit 2013] Marc Petit, Stéphane Lallée, J-D Boucher, Grégoire Pointeau, Pierrick Cheminade, Dimitri Ognibene, Eris Chinellato, Ugo Pattacini, Ilaria Gori, Uriel Martinez-Hernandez*et al. The coordinating role of language in real-time multimodal learning of cooperative tasks.* Autonomous Mental Development, IEEE, 2013. (Cited in page 10.)

[Rabideau 1999] Gregg Rabideau, Russell Knight, Steve Chien, Alex Fukunaga and Anita Govindjee. *Iterative repair planning for spacecraft operations using the ASPEN system.* In Artificial Intelligence, Robotics and Automation in Space, volume 440, page 99, 1999. (Cited in page 9.)

[Ramırez 2009] Miquel Ramırez and Hector Geffner. *Plan recognition as planning.* In Proceedings of the 21st international joint conference on Artifical intelligence. Morgan Kaufmann Publishers Inc, pages 1778–1783, 2009. (Cited in page 9.)

[Rich 2010] Charles Rich, Brett Ponsler, Aaron Holroyd and Candace L Sidner. *Recognizing engagement in human-robot interaction.* In Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on, pages 375–382. IEEE, 2010. (Cited in page 10.)

[Richardson 2007] Michael J Richardson, Kerry L Marsh, Robert W Isenhower, Justin RL Goodman and Richard C Schmidt. *Rocking together: Dynamics of intentional and unintentional interpersonal coordination.* Human movement science, vol. 26, no. 6, pages 867–891, 2007. (Cited in page 7.)

[Rizzolatti 2004] Giacomo Rizzolatti and Laila Craighero. *The mirror-neuron system.* Annu. Rev. Neurosci., vol. 27, pages 169–192, 2004. (Cited in page 6.)

[Ros 2010] Raquel Ros, Séverin Lemaignan, E Akin Sisbot, Rachid Alami, Jasmin Steinwender, Katharina Hamann and Felix Warneken. *Which one? grounding the referent based on efficient human-robot interaction.* In RO-MAN, 2010 IEEE, pages 570–575. IEEE, 2010. (Cited in page 10.)

[Roy 2000] Nicholas Roy, Joelle Pineau and Sebastian Thrun. *Spoken dialogue management using probabilistic reasoning.* In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, pages 93–100. Association for Computational Linguistics, 2000. (Cited in page 11.)

[Salam 2015] Hanan Salam and Mohamed Chetouani. *A multi-level context-based modeling of engagement in human-robot interaction.* In Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, volume 3, pages 1–6. IEEE, 2015. (Cited in page 10.)

[Sanghvi 2011] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan and Ana Paiva. *Automatic analysis of affective postures and body motion to detect engagement with a game companion.* In Human-Robot Interaction (HRI), 2011 6th ACM/IEEE International Conference on, pages 305–311. IEEE, 2011. (Cited in page 10.)

[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 1 and 6.)

[Sebanz 2009] Natalie Sebanz and Guenther Knoblich. *Prediction in joint action: What, when, and where.* Topics in Cognitive Science, vol. 1, no. 2, pages 353–367, 2009. (Cited in page 6.)

[Shah 2011] Julie Shah, James Wiken, Brian Williams and Cynthia Breazeal. *Improved human-robot team performance using chaski, a human-inspired plan execution system.* In Proceedings of the 6th international conference on Human-robot interaction, pages 29–36. ACM, 2011. (Cited in page 11.)

[Singla 2011] Parag Singla and Raymond J Mooney. *Abductive Markov Logic for Plan Recognition.* In AAAI, pages 1069–1075, 2011. (Cited in page 9.)

[Sisbot 2012] Emrah Akin Sisbot and Rachid Alami. *A human-aware manipulation planner.* IEEE Transactions on Robotics, vol. 28, no. 5, pages 1045–1057, 2012. (Cited in page 11.)

[Sorce 2015] Marwin Sorce, Grégoire Pointeau, Maxime Petit, Anne-Laure Mealier, Guillaume Gibert and Peter Ford Dominey. *Proof of concept for a user-centered system for sharing cooperative plan knowledge over extended periods and crew changes in space-flight operations.* In RO-MAN, IEEE, 2015. (Cited in page 10.)

[Talamadupula 2014] Kartik Talamadupula, Gordon Briggs, Tathagata Chakraborti, Matthias Scheutz and Subbarao Kambhampati. *Coordination in human-robot teams using mental modeling and plan recognition.* In IROS, IEEE/RSJ, 2014. (Cited in page 10.)

[Tenorth 2013] Moritz Tenorth and Michael Beetz. *KnowRob: A knowledge processing infrastructure for cognition-enabled robots.* The International Journal of Robotics Research, vol. 32, no. 5, pages 566–590, 2013. (Cited in page 16.)

[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. 05, pages 675–691, 2005. (Cited in pages 2 and 3.)

[Trafton 2013] Greg Trafton, Laura Hiatt, Anthony Harrison, Frank Tamborello, Sangeet Khemlani and Alan Schultz. *Act-r/e: An embodied cognitive architecture for human-robot interaction.* Journal of Human-Robot Interaction, vol. 2, no. 1, pages 30–55, 2013. (Cited in page 16.)

[Vesper 2010] Cordula Vesper, Stephen Butterfill, Günther Knoblich and Natalie Sebanz. *A minimal architecture for joint action.* Neural Networks, vol. 23, no. 8, pages 998–1003, 2010. (Cited in pages 8 and 11.)

[Walters 2007] Michael L Walters, Kerstin Dautenhahn, Sarah N Woods and Kheng Lee Koay. *Robotic etiquette: results from user studies involving a fetch and carry task.* In Proceedings of the ACM/IEEE international conference on Human-robot interaction, pages 317–324. ACM, 2007. (Cited in page 11.)