Talking On The Moon

Geert-Jan M. Kruijff[*]
Language Technology Lab
German Research Center for Artificial Intelligence (DFKI GmbH)
Saarbrücken, Germany
gj@dfki.de

Abstract

How effective, efficient and natural human-robot interaction can be depends for an important part on how meaning can be shared and construed between a human and a robot. The paper presents an initial investigation in the architectural functions that appear to underly embodied, situated understanding in joint action and interaction. The identified functions are combined into a concept system architecture. By interpreting these functions as requirements, the paper investigates dimensions for organizing groups of design decisions that affect how these requirements can be met in a concrete architecture. As a result, a space of architectural niches (conglomerations of functional requirements) arises, which each have different effects the possible deployment(s) of the robot. Reflecting on this given some salient aspects of the current state-of-the-art in HRI and cognitive systems, the paper raises open challenges for building robots for collaborative space exploration. (This paper presents work in progress.)

Introduction

Human-Robot Interaction (HRI) is a relatively novel field, and NASA's vision for space exploration presents it with a whole new set of challenges. These challenges not only concern how we can create embodied, interactive systems with whom we can collaborate in complex situations like planetary exploration. We also need to design systems that behave properly when deployed. Mistakes are expensive when they happen millions of miles away, and systems should never endanger the astronauts through unsafe behavior. It is thus important to understand the reasons for what a system can, and cannot, do. In principle these reasons arise from the design decisions we make when constructing the system – decisions that are based on requirements we need to meet, to solve a problem we have set ourselves.

My goal in this paper is to focus on a specific class of problems in this setting: Namely, that of *understanding*. Assuming the robot has a model of agency, with which it can reason about its own beliefs and intentions, and attribute such to others, how can a robot construe meaning out of its (embodied) experience of the environment, on which the robot's model of agency can then operate? Particularly, how may design decisions affect the kinds of understanding the robot is able to build, thus influencing its capabilities for collaborating with other agents?

It is a problem that we cannot ignore. The degree to which a robot can understand itself, the environment, and other agents has a fundamental impact on the effectiveness, efficiency and naturalness of human-robot interaction. This experience forms the basis on which the robot needs to act, not just in a reactive fashion, but pro-actively. It is in the very nature of collaborative action that requires the robot to have an explicit understanding not only of its own actions, but also of those of others, with the ability to predict and anticipate the combined effects these actions may have.

What makes this problem nontrivial is that the robot has multiple modalities for perceiving, manipulating, and moving in an environment that it may only partially know. Below, I will use available insights in cognitive systems to come to an abstract functional decomposition that indicates how this problem may be seen to arise from several basic issues: sensorimotoric *coordination*, *cross-modal content-association* between sensorimotoric modalities, and predictive *causality* to be able to anticipate effects of actions. Coordination between multiple sensorimotoric modalities enables the robot to form first of all a body-relative (egocentric) sense of space. Establishing coherence between the content across these modalities makes it possible to understand the relation between its own body and other objects in the environment in terms of action and perception. Causality connects this spatial understanding of action and perception with a temporal dimension, enabling the robot to predict and anticipate the possible effects of its own actions.

To understand how design decisions may influence the levels of understanding a robot may be capable of, I will proceed as follows. The first step is to provide the above-mentioned functional decomposition, from an assumed ability to conceive of other agents acting in the environment, back to the robot's own abilities to make (pro-active) sense of its experience. The purpose of this decomposition is to give a characterization of the different levels of interpreta-

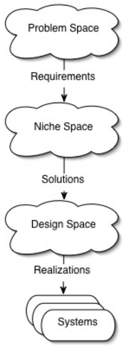
[*] The research reported of in this paper was supported by the EU FP6 IST Cognitive Systems Integrated project *Cognitive Systems for Cognitive Assistants* "CoSy" FP6-004250-IP. Thanks to project partners, particularly Aaron Sloman, for inspiring discussions, and the reviewers for challenging comments on an earlier version.

tion and abstraction involved in building up the kind of understanding we need for a collaborative robot. The second step is then to integrate this decomposition into a more general concept that includes agency. The point here is not to provide a model of agency, but to discuss how agency can act on the robot's understanding of its experience.

Subsequently, I outline how design decisions along four dimensions affect the aspects of understanding, and thus the shape that this concept can take in an actual design. These four dimensions are decisions about the *environment* in which the robot will be deployed ("Where?"), the *embodiment* of the robot ("What will we enable it to do? to observe?"), the *actions* of the robot ("What do we want it to be able to do?"), and the nature of the *interaction* between a robot and a human ("What do we want to talk about?"). Finally, I will reflect on the extent to which we are currently able to realize the concept.

With that, I hope to contribute in general to a deeper understanding of what it means to make collaborative robots – what levels of understanding we need if we want to walk and talk with them on the moon. More specifically, the paper points out that an important aspect of such understanding is that the robot is able to *associate* content from different modalities into a cross-modal understanding. What makes the presentation of this insight in this paper novel is its synthetic nature: Specific insights in the need for cross-modal content association in cognitive systems are combined into a larger picture, indicating the fundamental role cross-modal categorical systems play in building up spatiotemporally situated understanding. Design decisions affect the possible content of such categories, how the robot can use categories to mediate between modalities, what it can infer from this mediation, and to what extent that understanding can be used in communication about collaborative action.

The kind of analysis I propose in this paper is thus slightly different from the discussion that Fong & Nourbakhsh provide. Fong & Nourbakhsh identify various high-level functionalities that human-robot interaction should cater for. Following (Sloman 1994) I instead look at how the shape of such functionalities depends on the design decisions we make about environment, embodiment, and afforded interaction and action when thinking about how the system should be deployed. In this I follow a biology-inspired interpretation of (Sloman 1994): We need to establish the functional requirements (*niche*) that result from instantiating the four dimensions in particular ways (*problem*), and examine the possible systems that would meet these requirements (*design*).



Overview of the paper The paper first discusses the functional decomposition, and fits it into a concept that includes a notion of agency. Then, it is shown how design decisions along the dimensions of environment, embodiment & action, and interaction influence the nature of functionality in the concept, and thus the degree to which a robot can establish a situated understanding necessary for collaborating with other agents. The paper finishes with a brief reflection on the state-of-the-art, and the challenges we still face. The remainder of this section provides a short illustration of the kind of collaborative action and interaction we may be striving for – which will be referred back to at various points in the discussion.

Image you are on the moon, and you and your robot partner have been sent on a tactical mission to prospect for water repositories ... "Okay, Partner. Here we are, Basin X137", the robot says. Groaning, you climb out of the lunar rover – it's been a bumpy ride, and EVA suit and low gravity notwithstanding, you're damn stiff and happy to get going. "You alright?" "Yeah, I'm fine, thanks. Let's go." "Sure. I'll get the drill," the robot says, gesturing to the seriously heavy drill setup for collecting deep soil samples, "so if you could get the measuring devices, we're all set." You pick up the devices, look at the positioning information projected on your visor, and point in the direction the repositories are thought to be. "That way, down the ridge, into the gully."

As you walk to the ridge, you notice that the gully is too deep to just jump into. "We'll have to climb down. I'll go first," you say. "As you wish." You get over the ridge, carefully checking your footing. As you go down, all over sudden you loose your grip as a lava rock crumbles to dust in your glove, and you start sliding down faster than your physiology is appreciating. "Partner! Hold on!" "To what?," you scream. Looking up, you cannot help but wonder about the agility with which the robot is descending. 'Having six legs surely helps,' you think. 'So much for evolution.' Quickly, the robot stretches out an arm for you to grab, stopping you from sliding down, and helps you covering the last couple of meters in a more graceful way.

"Let me see your suit, Partner," the robot says. "Oh, stop nannying me!" "There is a scratch here, better fix that." The robot takes out some nanotech universal tape, puts it over the scratch, after which the tape quickly meshes with the suit. "There. Repaired." "Thanks. And thanks for grabbing me there." "No problem, Partner."

Pulling yourself and your pride together again, you get up. "Okay. Shall we set up the drill over there, by that rock?" you suggest. "You mean the big rock? Looks like some meteor impact debris," the robot says. "Sure does. Yeah, that's where I'd head." "Okay." Once you get to the rock, you hunch down to clear away smaller debris. "Could you set up the drill here? Then I'll get the measurements ready." "Okay, Partner. Could you briefly hold this?" The robot hands you the drill bit, while it puts the rig in place. 'Ha! And that despite 6 arms!' "Thanks, Partner," says the robot and holds out its hand for the bit. You hand it over, saying "Looks like 6 meters could be a good start." ...

An Initial Orientation

Before we can talk about what decisions we can take in designing a collaborative robot, we first of all need to understand what we can take these decisions *about*. What would make a collaborative robot tick? What is the kind of functionality that seems to be required, such that the robot not only understands what it does, and can do, but also what you can do, and how you could do things together?

In this section, I focus on what could be seen as a functional core that would enable the robot to establish such an understanding. To this end, I first construct a functional decomposition. This decomposition is inspired by cognitive systems, but its point is not to outline how these systems should develop, nor how the individual functions should be implemented. Rather, it focuses on characterizing functionality that appears to provide a necessary basis for enabling understanding: What does the robot need to be able to understand, first of all about itself, before it can understand others in such a way that it could collaboratively act and interact with them? I will argue that this ability ultimately rests on the abilities to *coordinate* different sensorimotoric modalities, establish *coherence* in content associations between these modalities, and employ a notion of predictive *causality* to derive predictions on the basis of which future events can be anticipated. On the basis of the resulting decomposition, I then discuss an initial concept. Whereas the decomposition presents the basis for understanding, the concept illustrates how this understanding could be used by a theory of agency for collaborative action and interaction.

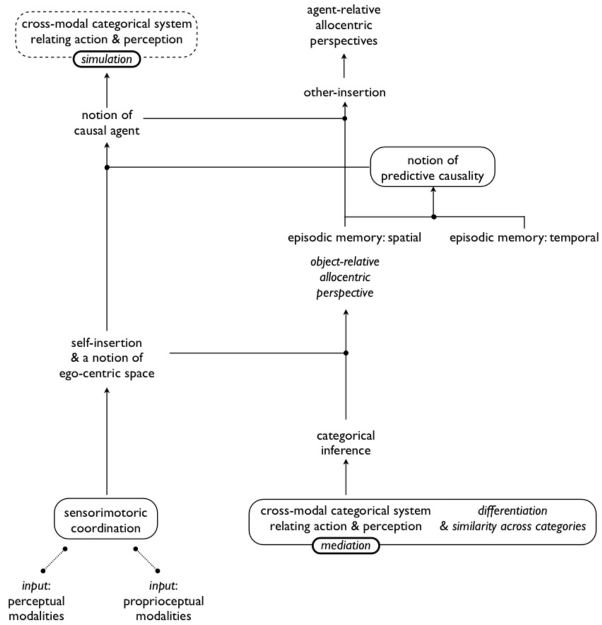


Figure 1: Functional decomposition

Architectural Functions

For the robot to collaboratively act and interact, it needs to be capable of understanding (and explicitly making) several abstractions. First, we have the abstraction from the ability to execute an act, to an understanding of the effects that would be yielded if the act were performed. This is essential for the robot to behave pro-actively, not just react. Second, the robot needs to be able to abstract from itself, as a causal agent, to the recognition of other agents as causal agents. This enables the robot to ascribe actions to other agents. But this is not enough. Third, the robot needs to be able to abstract from its ego-centric viewpoint, and consider the situation from the perspective of the other agent. Only this way will the robot be able to establish a common ground in understanding the situation, and consider the effects of the other's acts. The point of the functional decomposition I present here is to see from what underlying functionality these abstractions arise, to yield the understanding we are looking for in the robot. The abstraction steps are illustrated in Figure 1, given as arrows.

Assume that the robot is able to reason about agency, attributing beliefs and intentions to itself and to others; see also the discussion of the concept below. Ultimately, given the collaborative setting we are interested in, the robot thus needs to be able to apply its reasoning to how it understands the situatedness of the other agents. For this, the robot needs to be able to work out *agent-relative allocentric perspectives*. An agent-relative allocentric perspective is a perspective under which the robot can conceive of how the environment may appear from the viewpoint of another agent – which is normally different from the ego-centric viewpoint from which the robot itself perceives the situation. This affects how the robot can anticipate what that other agent may be capable of doing, in a concrete situation: allocentric perspectives are not only determined by the idea of taking on a different spatial reference point (Berthoz 2000) but also by functional aspects of the object or agent that the reference point is transfered to, cf. e.g. (Coventry & Garrod 2004; Carlson & Kenny 2005) for the object-relative view.

For the robot to construct such a perspective, it needs to recognize another agent as a *possible* reference point for such a perspective. We can call this *other-insertion*, the explicit recognition of other agents in the environment, derived from *self-insertion* as the recognition of oneself as a causal agent (Philipona, O'Regan, & Nadal 2003; Philipona *et al.* 2004); see also below. In turn this recognition relies on the robot *having* a notion of causal agent.

Based on causal agency the robot is able to understand that agents act and that these actions have effects – cf. the notion of intentional stance (Dennett 1987). Other-insertion makes it possible to see that there may be others besides the robot that can affect the environment, and agent-relative allocentric perspectives enable the robot to see how that may happen. These three capabilities lead to an understanding that is fundamental to collaborative action. Without the ability to develop (agent-relative) allocentric perspectives, the robot would essentially be an autistic system (cf. e.g. (Berthoz 2000; Frith & de Vignemont fc)). The robot would be restricted in the extent to which it could collaborate as it would only have a limited conception of other agents beyond itself. On the other hand, lacking the ability to attribute causal agency to others would present a control problem to the robot. The robot would understand that an agent did something, but not who – which could degenerate into an artificial equivalent of schizophrenia; cf. also (Proust 2000).

Being able to see others as causal agents presumes that the robot is able to see itself as a causal agent operating in the environment. On the one hand, this requires the robot to be capable of *self-insertion*, seeing itself as acting in the environment. Self-insertion is based in a notion of ego-centric space, which derives from the ability of the robot to *coordinate* its sensorimotoric modalities, relating action and perception. See for example (Berthoz 2000; Philipona, O'Regan, & Nadal 2003; Philipona *et al.* 2004).

On the other hand, the robot should not only be able to perceive that it acts in the environment, but also that its actions have effects – and how these effects could be predicted, and thus anticipated. For this, we need a notion of *predictive causality*; cf. e.g. (Sloman 2005). We need a notion of causality that is stronger than just presenting empirical correlation: It needs to provide not only the possibility to generalize over previously seen situations, but also to recombine knowledge to predict the effects in novel situations.

This presents requirements on the organization of the semantic knowledge of the robot: it should not only make clear how semantic categories differ, but also to what extent they are similar – what features, what parts – as similarity can be used as the basis for generalization to new instances. This is an issue which I return to below. A more immediate requirement is that the robot has a notion of *events*, which is based on a reflection in *episodic memories* dealing with the *temporal* and *spatial* structure of events.

Although temporal episodic memory is an important aspect here, I would like to focus on the issues concerning spatial episodic memory. A fundamental characteristic of spatial episodic memory is its *allocentric* representation of objects, and situations ("cognitive maps"); cf. e.g. (Burgess, Maguire, & O'Keefe 2002) and the citations therein. This represents an abstraction from the ego-centric viewpoint of the robot, arguably based on the functionality of the objects onto which the reference point is transfered; cf. e.g. the discussions in (Grush 2001; Coventry & Garrod 2004; Carlson & Kenny 2005).

To establish this functionality, we need a *cross-modal categorical system* that enables us to derive this functionality from categories that model how the actions on such objects yield particular effects. Core to the notion of this categorical system is thus a sense of affordance, cf. e.g. (Barsalou 1999; Glenberg 1997; Glenberg & Kaschak 2002; Borghi 2005). At the most basic level, this results in a specification of how to interact with an object – cf. the notion of μ-affordances (Ellis & Tucker 2000).

This accords a fundamental role to a cross-modal categorical system in forming understanding. Sensorimotoric coordination, and bringing about a coherence in content-association between modalities, need to go hand in hand in this process. The coordination enables categorization, and in particular the combination of perception and action. On the other hand, associating the content in this coordination enables the categorical inference which provides the basis for completing the perceptual input on the basis of previous experience, and helps the sensorimotoric modalities to mutually constrain what input needs to be considered (Berthoz 2000). Furthermore, categorical inference over categories enable the prediction, and thus the anticipation, of possible effects of actions – cf. also the notion of simulation in (Barsalou 1999). The cross-modal nature of categories is thus crucial to bring about a sense of affordance (be that μ-affordance, or Norman's notion). Furthermore, coupled to a spatiotemporal perspective, it enables us to abstract away from the current situation, or our own ego-centric experience, and generalize towards future situations or other perspectives.

Concept

Above I discussed how we can conceive of what appears to be necessary for a robot to build up an understanding of the environment it is situated in. This understanding is reflected in an spatiotemporal episodic understanding of the current situation, with the possibility to entertain different perspectives on this situation and possible future situations on the basis of inferences over a cross-modal categorical system that enable the robot to predict the situated effects of actions on objects.

Assuming we have a model of agency suitable for modeling collaborative action and interaction (e.g. the SharedPlan approach of (Grosz & Kraus 1996; Lochbaum 1998)), the question is how can employ the above understanding in our model of agency. To this end, we need to make it possible to mediate between the representations we use in our model of agency, and the deeper levels of semantic and spatiotemporal episodic understanding. This requires that we can map the ontological distinctions, made in our content models and plans for action & interaction, to the categorical information presented in episodic memory and the categorical system. The resulting connection between plans, and episodic and semantic understanding would make it possible to establish the situated feasibility of plans, and establish a common ground between the robot and other (communication) partners. Conversely, as I will point out below, establishing a connection from episodic understanding to the model of agency provides interesting possibilities for active situational awareness, i.e. the ability to adapt plans on the basis of an understanding how changes in the environment affect intended actions.

Design Dimensions

When we design a robot, there are usual several dimensions along which to consider the design. One, we have the environment in which it is to be deployed. What we consider this environment to be like influences what we believe there to be for the robot to perceive. Ultimately, this affects what we enable the robot to know and reason about, thus determining the "metaphysics" of the system. Second, we have the robot's embodiment. This plays a crucial role in what we want to robot to perceive and do, and how. Third, there is the issue of interaction – the decisions we make about how the robot can communicate with others. Because communication is for an important part *about* the robot's understanding of the environment it is situated in, the decisions we make on the first two issues can have an important impact on the robot's abilities to (sensibly) communicate.

Environment

An inherent aspect of embodied interaction is that the action and interaction take place in, and refer to, the reality in which the agents are situated.[1] This reality is external to the agents, and its character might be independent of how the agents conceive of the reality surrounding them.

This raises issues in what *perceivables* the reality offers. I understand a perceivable to be data about the current state of the environment. The nature of a perceivable is circumscribed by at least the following aspects:

- *Existence*: Does the perceivable regard an endurant ("object") or a perdurant ("process")?
- *Intro/extro-spective nature*: Does the perceivable provide an introspective window into the state of the environment, or only an extrospective characterization?
- *Certainty*: How certain is the data that the perceivable provides?

Another issue is whether the reality *changes*, or whether it remains in a singular state. Clearly, this has an influence on the nature of perceivables – processes in a static environment can only be considered as immutable states.

The decisions we make about perceivables have, as I already pointed out, an important effect on the "metaphysics" of what the robot will be able to know and reason with. For example, consider the possible effects of what we consider existence to be like. If we would decide that the only relevant aspects of the environment regard objects, not processes, then temporal episodic memory will be restricted to states, and the robot's notion of change will be discrete. As a result, causes will appear "instantaneously." This may be easier to model than (semi-)continuous processes, but it raises the question to what extent the robot will then be able to observe that a change is being brought about by other agents – and thus, whether the robot is will be capable of other-insertion. The other two issues, the intro/extrospective nature of perceivables, and certainty, have an effect on the degree to which the robot will be able to know something (independent of the characterization of its own perceptual modalities), and establish the possible effects of action on it.

Rather than going into the philosophical underpinnings of the above view of reality (and any notion of "ground truth" it may give rise to), I would like to illustrate the parametrization of perceivability and changability of the environment on some familiar examples.

The simplest form of 'reality' is the virtual blocks world with some objects. We can model such a world as a database, containing full descriptions of the objects and their locations in the world. In such a world, perceivables only concern endurants, whose nature we can completely introspect by querying the database. With a real blocks world we loose introspection of the state of objects, and introduce potential uncertainty of what state an object is in.

As another example, let us take the office environment, a popular research environment in mobile robotics – preferably "not necessarily" populated with humans. These environments provide an interesting extension over block worlds, as they usually include at least changing states such as open/semi-open/closed doors. Because these states are analog, uncertainty is extended to processes as well. The actual extent to which perceivables are (un)certain depends on the instrumentation of the environment, and whether ot makes perceivables available.

Our story illustrates several aspects of complex environments we are likely to encounter in places like the moon. The environment includes objects, both known (e.g. the drill and its parts) and unknown (the rock), and processes both semi-controlled (e.g. drilling) and uncontrolled (e.g. breaking, falling stones).

Embodiment & action

The literature on cognitive systems provides various notions of embodiment, cf. (Ziemke 2001). For my current purposes, I abstract away from biological concerns with embodiment, and focus on three aspects of embodiment that Ziemke notes: structural coupling between the agent and the environment, the nature of the physical embodiment of the agent, and the historical (or episodic) aspect of embodiment.

- *Observables*: Given the structural coupling between the agent and the environment, what perceivables can the agent interpret as observables on which it can deliberate behavior?[2] Given that observables are perceptions the agent is aware of, what levels of abstraction do observables represent?

Design decisions on the nature of observables have a fundamental impact on the formation of categories in the robot's categorical system. At the most basic level, there are the issues regarding "perceptual grounding" of these categories, i.e. the primitive features for e.g. objects (Roy to appear; Oates 2003), actions (Baily 1997; Naranayan 1997; 1999). Possibilities for differentiation between categories on the basis of these features, and their (re)combinability, then determine the levels of abstraction the robot can establish in a categorical system – affecting the categorical inferences the robot can make, and what kind of coupling between perception and action can be established.

Observables do not need to correspond to individual perceivables, as illustrated by e.g. localization in the Pygmalion system described in (Siegwart & Nourbakhsh 2004)(p.238ff.). The system operates in a reality in which contains objects that provides perceivables which we assume we can detect with a laser rangefinder.[3] It bases its localization and navigation behavior on observables that are *lines*, i.e. abstractions over collections of individual point range measurements (the perceivables).

- *Physical embodiment*: What types of locomotion and manipulation does the agent's physical embodiment enable?

Depending on its physical embodiment, the agent can usually perform various types of actions, ranging from simple movement to complex manipulation. One of the first questions to ask in this context is how actions are determined

[1] Modulo multiple spatial ranges.

[2] As Werner Heisenberg put it, "[W]hat we observe is not nature in itself but nature exposed to our method of questioning."

[3] This is a sometimes critical assumption, as laser cannot detect glass objects.

– are they reactive? Are they planned? Several hybrid architectures have been proposed, mixing reaction and planning – but then, how are these layers connected?

These questions have been asked before, cf. for example (Brooks 1986; 1991) and the discussion of hybrid architectures in (Arkin 1998). Looking at action from the viewpoint of embodiment puts them in a new light, though. It is not enough to know *that* an agent can perform actions. We also want to know *why* the agent's embodiment enables these actions, such that we can combine this with information from the environment to understand *when* actions can be performed – either by the agent, by others, or jointly.

- *Layers of processing*: At what layers are the agent's actions determined? How are these layers connected?
- *Connectivity between sensorimotoric modalities*: At what layers are motoric and perceptual modalities connected? How are they connected?

Layered processing usually follows an idea of functional abstraction. This idea is familiar from neuroscience, as it is a fundamental organizational principle in the brain, cf. (Kandel, Schwartz, & Jessell 1991). Design decisions about connectivity between sensorimotoric modalities affect first of all the possibilities for establishing coordination between these modalities. How this coordination can then be coupled to a categorical understanding of perception and action depends on the correspondence between the levels of abstraction at which we establish this connectivity, and the levels of description in the cross-modal categorical system.

Furthermore, the (in)ability to coordinate sensorimotoric modalities can have a profound effect on the robot's self-insertion, and -at a more abstract level- its ability to conceive of the presence of other agents in the environment.

In the story, the robot combines reactive, planned, and joint action with active perception of affordances for its own actions and those of the human partner. While climbing down, the robot is aware of what rocks afford stable stepping. Reactive behaviors handle the actual stepping, guiding by planned paths. Active situational awareness makes the robot rapidly change its path, to aim for the falling human partner. It quickly plans a path that, given the sliding human, will bring the robot into a position that affords to human to grab one of the robot's arms, and that affords the robot a stable position to safely stop the human from falling. Other examples of joint action include the carrying and deployment of the drilling rig.

Embodiment and action, by its very nature, implies an *active presence* of the agent in the environment. If we relate this to the previous discussion about decisions on the nature of the environment, more issues arise:

- *Active situational awareness*: How aware is the agent of changes in the environment? How can the agent be attentive to change in the environment?
- *Experience and learning*: To what extend, and how, is the agent able to learn about the environment? How plastic is the acquired experience? To what extend is the agent able to reconsider and modify acquired experience?

For agents operating in dynamic environments, it is important to be aware of how the environment is changing as changes may require the agent to adapt its current and future behavior. This crucially relies on attentional mechanisms (Rensink, O'Regan, & Clark 1997): People tend to be blind to change unless they pay attention to it (O'Regan to appear). The combination of sensorimotoric coordination and a cross-modal categorical system may prove to be crucial here, as categories may present associations between modalities that enable selective attention.

The decisions we make about the robot's levels of awareness to change, the flexibility in dealing with change, and the nature of experience & learning all relate to Ziemke's episodic aspect – i.e. to the nature of the episodic structure in spatial and temporal episodic memories. Underlying these issues are, of course, the deeper decisions on what the robot can in principle know about in the environment, and to what extent the robot is able to interpret its own situatedness. All these decisions conspire to delineate the nature of episodic structure, which in turn affects on how the robot can interact: As I already indicated earlier how processing collaborative action and interaction relies on the possibility to reflect its content in these episodic memories.

Interaction

With interaction, we can attempt to close the circle and bring all the dimensions together. I understand interaction to be *communication* between agents, and I want to set interaction apart from action. In communicative interaction, agents exchange information, to form a common ground on which they can establish an intersubjective interpretation of the information. Interaction thus has an informing and coordinating function, allowing for organization that can concern not only the present and the past, but also future events. Interaction thus enables the agent to go beyond the confines of personal experience and affordance.

In embodied interaction, an agent interprets communication against the situated environments, given the embodiment characteristics and the intentions of the involved agents; (Dourish 2001). The resulting notion of meaning, Dourish argues, corresponds to Norman's notion of affordance. This first of all raises issues regarding the situated nature of interaction.

- *Situatedness*: To what extend is the agent able to ground the meaning of communication in the situatedness of the interaction?
- *Experiental grounding*: To what extend is the agent able to ground the meaning of communication in its own experience, and the (inferred/projected) experience of others?

Both issues on the connection from processing collaborative action and interaction, to episodic memory and the cross-modal categorical system. Experiental grounding relies on the degree to which we can interpret (communication about) action plans in terms of (potential) situated action – following (Glenberg 1997), how can we connect the plans we talk about with how we can actually *mesh* the affordances provided by the objects to which the physical actions are to be applied? Whereas plans are abstract, experiental grounding provides the possibility to consider the executability of these plans in a concrete situated context. A similar issue concerns situatedness: to what degree are we able to process references to the situated context, on the basis of the robot's own understanding of the spatiotemporal aspects of the situation?

But, we also have an episodic connection in the other direction:

- *Flexibility and adaptability*: To what extend is the agent able to be flexible, and adaptive, in how it communicates with other agents?
- *Interactive affordances*: How do the interactive skills of the agent afford other agents to use particular interactive skills?

The first issue concerns design decisions on how perceived changes in the environment can actually trigger changes in plans (and content models) for action and interaction – without these triggers, continuous top-down verification and reinterpretation may be a consequence.

The story provides an illustration of how effective, efficient and natural collaborative interaction naturally *involves all of the above. Humans typically have a qualitative, context-sensitive understanding of the spatial organization of an environment, and the robot is able to seamlessly relate such an understanding with its own quantitative perception to process references to the rock and to "here".*

Furthermore, the robot understands that the drill rig is too heavy for its human partner, but that neither the measurement devices nor the drill bit pose a particular challenge to the human's physical strength. Using this understanding, it can take the initiative in the dialogue, and quickly coordinate with the human how to carry out the plan of getting the drilling rig in place to collect samples. And precisely because the robot has these interaction skills, it affords the human to use similar skills a bit later to coordinate where to start drilling.

Finally, the robot's ability to episodically connect action and interaction makes it possible for these different dialogues for coordinating actions, and the joint actions themselves, to blend into each other. There is no need for the human to re-instruct the robot because the situations changed between discussing and refining the plans again, and carrying them out.

Requirements arise from coupling the dimensions

Before designing a system, we need to determine in what kinds of environments it would be deployed, and what it should be able to do there. This results in a set of design decisions concerning the nature of the environment, embodiment & action, and interaction. They delineate the system and its deployment, and thus give rise to functional requirements for the system design to meet. Naturally, I do not assume that moving between design, assumptions and requirements is a linear process –it is a gradual process of refinement– nor that a collection of requirements (a niche) determines a unique design. We are exploring spaces of possibilities – problem spaces, niche spaces, and design spaces (Sloman 1994).

First, consider a simple robot for planetary exploration, e.g. Brook's Hannibal or Attila. In its simplest form, it should just be able to move in rocky environments. For such a robot, we can assume the environment consists of unknown objects (e.g. rock) which can only be extrospectively perceived to some degree of certainty. Furthermore, the robot's embodiment provides means for stable and robust movement in uneven terrain, and sensors to perceive the immediate environment. Sensoric perception directly triggers reactive behaviors – it is debatable whether it has any observables, as there is no 'conscious' observation.

The robot thus provides some action, but no notion of affordances, nor any means for interaction. What if we *would* want to add interaction? For example, assume that there may be metal deposits, and that this is the only thing the robot is able deliberatively observe, using a metal detector. The possibilities for interaction in such a case are extremely limited. The robot lacks positioning information, and is thus not able to deal with situatedness. It can only tell whether it observes a metal deposit or not, and -lacking any notion of embodiment or presence, or episodic connection between its reactive moves and the observations it can communicate-there is no room for flexibility or adaptivity in communication. Another agent is restricted to two skills: listen to what the robot can say ("yes/no metal deposit"), or ask whether it observes a metal deposit.

A more advanced case could be a delivery robot that operates indoors at a moon base. We can perhaps assume that such an environment is highly structured, and well-instrumented. All objects and processes are known, and are instrumented to provide more information about their properties and state. Agents carry nametags that identify them, and provide exact information about their position. The robot's embodiment should enable it to move indoors, to carry objects, and to actively sense the instrumented environment. The robot only needs a binary notion of affordance (there/not there) – let us assume that all possible affordances can be known in a structured and instrumented environment. These assumptions have important consequences for the nature of action and interaction. The situations in which the robot will find itself are exhaustively predictable, and both actions/plans and interaction can thus be scripted beforehand. Within the limits of the scripted interaction, the human can use its own skills – but not beyond – and refer to the situation insofar as the robot can obtain information about the situated context through actively sensing instrumentation. Similarly, episodic connections can be established beforehand, and made such that interaction state equals action state.

The resulting robot will be able to carry out its task, and interact with other agents. Examples of such robots are (Ishiguro *et al.* 2001; *et al* 2004,); the latter explicitly make the assumption about interaction state equalling action state. Such robots are, however, only well-behaved in an environment where structure and available instrumentation can guide the robot. If random events occur, the robot may be able to respond in a local fashion (e.g. object avoidance of unexpected, unknown objects) but the question is what in such a case the effects would be on the global behavior of the robot, and whether it could gracefully overcome an erroneous state transition. Another aspect concerns the level of collaboration these robots are able to provide: joint actions can only be coordinated or adapted as long as this corresponds to predefined plans.

Clearly, these are not the robots we are thinking of when imagining fully-fletched robotic partners with whom we can work together. If we would want a robot as in the story, then we would need a design that comes close to the concept I discussed above.

Where In Design Space Are We?

How far are we when it comes to meeting these requirements? The above discussion should make clear that the effectiveness, efficiency and naturalness of collaboration depends on how human and robot can meaningfully act and interaction. And that, in the end, is determined for each agent by the afforded active and interactive skills, as well as its sense of the environment, of itself, and of the others involved.

Current research in HRI, and related fields such as computational linguistics, cognitive systems, artificial intelligence, and robotics, does not yet enable us to fully meet the requirements. Time and space prevent me to give a representative overview of the current state-of-the-art, to do justice to all the ongoing work that is relevant. A general observation we can make however is that research regarding cognitive systems usually follows either a bottom-up approach, or a top-down approach. And the problems in addressing the requirements arise precisely from the chasm between the complementary perspectives that these approaches take.

In a bottom-up approach, a particular architecture is designed, and it is investigated what behaviors emerge from the architecture in action. Typically, these architectures have a close structural coupling between the system and the perceivable reality, and are often sub-symbolic in nature. For example, work such as (Philipona, O'Regan, & Nadal 2003; Philipona *et al.* 2004) discusses how embodiment and sensorimotoric interact to discover a notion of space in which the embodied agent finds itself, fundamental to the notion of self-insertion. Other work has looked at the perceptual grounding of the meaning of objects and their properties (Roy 2002; Oates 2003), the perceptual grounding of action verbs (Baily 1997; Naranayan 1999)

In a top-down approach, we usually start with a set of (abstract) behaviors, and design an architecture around it. This approach is often adopted in symbolic approaches, modularizing different deliberative processes such as action planning or language understanding. This is also the predominant strategy in HRI, reflected in finite-state based systems such as RoboVie (Ishiguro *et al.* 2001) or Biron (*et al* 2004,), or more flexible approaches such as GODOT (Bos, Klein, & Oka 2003) and explicitly collaborative systems like Mel (Sidner *et al.* 2004) or the architectures for collaborative control discussed in (Fong, Thorpe, & Baur 2003; Bruemmer & Walton 2003).

The problem is that both types of approaches are addressing the problem of designing cognitive systems from different ends, without meeting in the middle. The bottom-up approaches have a close structural coupling, but often lack the connection to higher-level deliberative processes. With top-down approaches it is often the other way around.

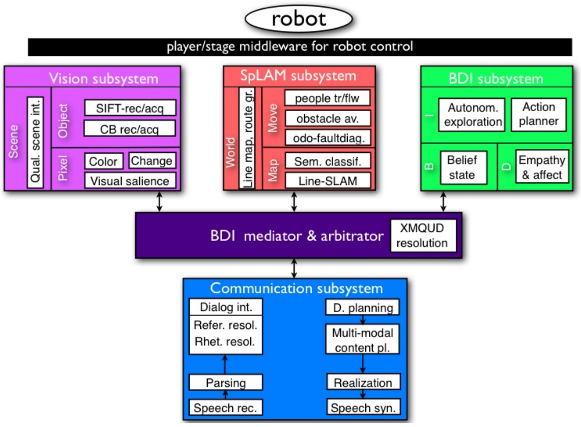


Figure 2: The CoSy architecture

In the CoSy project, we are investigating the requirements and possible designs for cognitive systems that operate in dynamic environments. Figure 2 presents the current architecture, which combines subsystems for visual observation, spatial localization & mapping, action-related BDI, and communication. Like most current architectures, it is distributed (using OAA and CORBA).

The architecture enables a close coupling between observation and deliberative processing through multiple levels of ontology-based mediation. We use ontologies to relate information across different modalities, enabling both early and late fusion, and to relate quantitative and qualitative interpretations of observations. We are currently exploring how we can base these ontologies more closely on aspects of the embodiment. The architecture also includes a rudimentary notion of physical presence (tracking, sphere of influence, odometric fault diagnosis), which we hope to extend towards a notion of presence that includes both self-insertion and other-insertion, embodied and related to recognized affordances. Active situational awareness is currently provided in a purely push-based fashion, modulated by local visual attention: based on recognized changes in the visual scene, the new observations are pushed to a module in which we provide a qualitative interpretation of the visually situated context. Through global fusion (in the belief state), the architecture can relate representations of e.g. discourse referents in the communication subsystem to their -possibly changed- observable properties. This is the basic mechanism for processing of situated communication. The communication subsystem provides a planning-based approach to dialogue processing, enabling basic strategies for mixed-initiative collaborative dialogue.

Currently, we are exploring resulting system in scenarios that involve among others human-assisted exploration of indoor environments ("human-augmented mapping"). The principle mode is that the human guides the robot around, and explains the spatial organization of the environment and relevant landmarks. The human only provides the high level information, though. The robot is capable of autonomous exploration, in which it can automatically create maps of the environment, semantically classify areas. Driven by a desire to discover, the robot can ask the human to provide more information (e.g. "What is behind this door?"), and to clarify uncertainties ("Is there a door here?").

Back To The Future

The CoSy architecture puts a strong emphasis on mediation between different levels of processing. This mediation enables us to fuse and share information across these levels, across modalities, and across quantitative or qualitative interpretations. Arguably, this provides a closer coupling between observational and deliberative processes than is available in other HRI architectures.

The architecture still differs from the often strongly associative architectures found in bottom-up, subsymbolic approaches to cognitive systems. And this is a problem, a problem it shares with other HRI architectures. With the distributed nature of our top-down designed architectures comes inevitably a modularization of information processing. The problem with modularization is that it encapsulates not only information processing, but also information and its interpretation, disconnecting it from the source it originated from. This presents a serious problem for operating in dynamic environments. Observing change requires awareness of what can change, and –through attention– observing when something changes, and how. This requires a mixture of bottom-up informing and top-down modulating and guiding, in an architecture that relates content rather than insulated processes.

I would like to advance the suggestion that one way to overcome this problem would be to move to content-addressable, multi-level architectures for embodied cognitive systems. The idea behind these architectures is to combine the ideas of associative content-addressable memories with multi-level functional abstraction/mediation grounded in embodiment, and with information processes that operate concurrently on addressable shared content in these memories. This would provide us with novel ways in which we could establish links between perceptual input and deliberative processes. Perceptual input could more immediately activate associated actions and interaction strategies. Furthermore, perception could be more directly guided and primed by attentional mechanisms that are connected to deliberative processes and the episodic content they are associated with.

Most fundamentally, these architectures would enable us to keep "the interpretant in the loop." As Dourish and Norman indicate, and before them people like Peirce, meaning is not just an issue of embodiment and environment. It crucially depends on the perspective that intentionality sheds on how the embodied observations of the environment are interpreted in context. My contention here is that only a mediated structural coupling between perception and addressable content maintained in episodic memories, accessed by deliberative processes, can establish this perspective. Provided mediation is based in features determined by the agent's embodiment, and the architecture provides the means for mental "simulation" in the form of e.g. mirror states, this structural coupling provides the basis for the determination of affordances, presence, and (intentional) perspective.

Conclusions

The effectiveness, efficiency and naturalness of collaboration depends on how human and robot can meaningfully act and interaction – and that, in the end, is determined for each agent by its sense of the environment, of itself, and of the others involved. The NASA vision for space exploration presents with novel challenges for HRI that take us well beyond the current state-of-the-art. In this paper, I explored what appear to be necessary functions for establishing an understanding that underlies succesfull collaboration, and discussed how design decisions may affect the extent to which we can achieve that functionality, and thus the extent to which a robot is capable of effective, efficient and natural collaboration. I illustrated how simple as well as more complex robots for lunar missions could be characterized this way – and particularly, how decisions affected how we could (not) interact with these systems. I ended with a brief look at current state-of-the-art systems for HRI. I argued that we need to overcome the gap that still exists in the structural coupling of perception and deliberative processes, if we want to build systems that can succesfully act and interact with other agents in dynamic environments. One possible way to do this would be to base architectural design on addressing shared content rather than on connecting encapsulated information processes. This would enable the robot to form a level of understanding that is reflected in an spatiotemporal episodic understanding of the current situation, with the possibility to entertain different perspectives on this situation and possible future situations on the basis of inferences over a cross-modal categorical system that enable the robot to predict the situated effects of actions on objects.

References

Arkin, R. C. 1998. *Behavior-Based Robotics*. Intelligent Robots and Autonomous Agents. Cambridge MA: MIT Press.

Baily, D. 1997. *When Push Comes To Shove: A Computational Model of the Role of Motor Control in the Acquisition of Action Verbs*. Ph.D. Dissertation, Computer Science Division, University of California at Berkeley.

Barsalou, L. W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences* 22:577–660.

Berthoz, A. 2000. *The Brain's Sense Of Movement*. Perspectives in Cognitive Neuroscience. Cambridge MA: Harvard University Press.

Borghi, A. M. 2005. Object concepts and action. In Pecher and Zwaan (2005). 8–34.

Bos, J.; Klein, E.; and Oka, T. 2003. Meaningful conversation with a mobile robot. In *Proceedings of the Research Note Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*.

Brooks, R. A. 1986. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation* 2:14–23.

Brooks, R. A. 1991. Intelligence without representation. *Artificial Intelligence* 47:139–159.

Bruemmer, D., and Walton, M. 2003. Collaborative tools for mixed teams of humans and robots. In *Proc. of the Workshop on Multi-Robot Systems*.

Burgess, N.; Maguire, E.; and O'Keefe, J. 2002. The human hippocampus and spatial and episodic memory. *Neuron* 35:625–641.

Carlson, L., and Kenny, R. 2005. Constraints on spatial language comprehension: Function and geometry. In Pecher and Zwaan (2005). 35–64.

Coventry, K., and Garrod, S. 2004. *Saying, Seeing and Acting. The Psychological Semantics of Spatial Prepositions*. Essays in Cognitive Psychology Series. Lawrence Erlbaum Associates.

Dennett, D. 1987. *The Intentional Stance*.

Dourish, P. 2001. *Where The Action Is: The Foundations Of Embodied Interaction*. Cambridge MA: MIT Press.

Ellis, R., and Tucker, M. 2000. Micro-affordance: The potentiation of components of action by seen objects. *British Journal of Psychology* 91:451–471.

et al, A. H. 2004. Biron - the Bielefeld robot companion. In Prassler, E.; Lawitzky, G.; Fiorini, P.; and Haegele, M., eds., *Proc. Int. Workshop on Advances in Service Robotics*. Stuttgart, Germany: Fraunhofer IRB Verlag. 27–32.

Fong, T., and Nourbakhsh, I. 2005. Interaction challenges in human-robot space exploration. *ACM Interactions* 12(2):42–45.

Fong, T.; Thorpe, C.; and Baur, C. 2003. Robot, asker of questions. *Robotics and Autonomous Systems* 42:235–243.

Frith, U., and de Vignemont, F. f.c. Egocentrism, allocentrism and asperger syndrome. *Consciousness and Cognition*.

Glenberg, A., and Kaschak, M. 2002. Grounding language in action. *Psychonomic Bulletin & Review* 9(3):558–565.

Glenberg, A. M. 1997. What memory is for. *Behavioral and Brain Sciences* 20:1–55.

Grosz, B. J., and Kraus, S. 1996. Collaborative plans for complex group action. *Artificial Intelligence* 86(2):269–357.

Grush, R. 2001. Self, world and space: on the meaning and mechanisms of egocentric and allocentric spatial representation. *Brain and Mind* 1(1):59–92.

Ishiguro, H.; Ono, T.; Imai, M.; Maeda, T.; Kanda, T.; and Nakatsu, R. 2001. Robovie: an interactive humanoid robot. *Int. J. Industrial Robotics* 28(6):498–503.

Kandel, E.; Schwartz, J.; and Jessell, T., eds. 1991. *Principles of Neural Science*. New York NY: Elsevier. 3rd edition.

Lochbaum, K. E. 1998. A collaborative planning model of intentional structure. *Computational Linguistics* 24(4):525–572.

Naranayan, S. 1997. *KARMA: Knowledge-Based Active Representations For Metaphor and Aspect*. Ph.D. Dissertation, Computer Science Division, University of California at Berkeley.

Naranayan, S. 1999. Moving right along: A computational model of metaphoric reasoning about events. In *Proceedings of the National Conference on Artificial Intelligence AAAI'99*.

Norman, D. 1994. *The Psychology of Everyday Things*. New York NY: Basic Books.

Oates, T. 2003. Grounding word meanings in sensor data: Dealing with referential uncertainty. In *Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data*, 62–69.

O'Regan, J. K. to appear. Change blindness. In *Encyclopedia of Cognitive Science*. Nature Publishing group.

Pecher, D., and Zwaan, R. A., eds. 2005. *Grounding Cognition: The role of perception and action in memory, language, and thinking*.

Philipona, D.; O'Regan, J. K.; Nadal, J.-P.; and Coenen, O. J.-M. 2004. Perception of the structure of the physical world using unknown sensors and effectors. *Advances in Neural Information Processing Systems* 15.

Philipona, D.; O'Regan, J. K.; and Nadal, J.-P. 2003. Is there something out there ? Inferring space from sensorimotor dependencies. *Neural Computation* 15(9).

Proust, J. 2000. Awareness of agency : Three levels of analysis. In Metzinger, T., ed., *The Neural Correlates of Consciousness*. Cambridge MA: The MIT Press. 307–324.

Rensink, R.; O'Regan, J.; and Clark, J. 1997. To see or not to see: the need for attention to perceive changes in scenes. *Psychological Science* 8:368–373.

Roy, D. 2002. Learning words and syntax for a scene description task. *Computer Speech and Language* 16(3).

Roy, D. to appear. Semiotic schemas: A framework for grounding language in the action and perception. *Artificial Intelligence*.

Sidner, C.; Kidd, C.; Lee, C.; and Lesh, N. 2004. Where to look: A study of human-robot engagement. In *Proceedings of the ACM International Conference on Intelligent User Interfaces (IUI)*, 78–84.

Siegwart, R., and Nourbakhsh, I. R. 2004. *Introduction to Autonomous Mobile Robots*. Intelligent Robotics and Autonomous Agents. MIT Press.

Sloman, A. 1994. Explorations in design space. In *Proceedings of the 11th European Conference on Artificial Intelligence (ECAI'94)*, 578–582.

Sloman, A. 2005. Two views of child as scientist: Humean and kantian. Presentation to Language and Cognition Seminar, School of Psychology, Birmingham University. Available from [http://www.cs.bham.ac.uk/research/projects/cosy/presentations/child-as-scientist.pdf].

Ziemke, T. 2001. Are robots embodied? In *Proceedings of the First International Workshop on Epigenetic Robotics — Modeling Cognitive Development in Robotic Systems*, volume 85 of *Lund University Cognitive Studies*, 75–83.