# Constructing Human-Robot Interaction with Standard Cognitive Architecture

Kristiina Jokinen

AI Research Center, AIST Tokyo Waterfront, Japan
kristiina.jokinen@aist.go.jp

**Abstract.** This paper discusses how to extend cognitive models with an explicit interaction model. The work is based on the Standard Model of Cognitive Architecture which is extended by an explicit model for (spoken) interactions following the Constructive Dialogue Modelling (CDM) approach. The goal is to study how to integrate a cognitively appropriate framework into an architecture which allows smooth communication in human-robot interactions, and the starting point is to model construction of shared understanding of the dialogue context and the partner's intentions. Implementation of conversational interaction is considered important in the context of social robotics which aim to understand and respond to the user's needs and affective state. The paper describes integration of the architectures but not experimental work towards this goal.

**Keywords:** Human-robot interaction, cognitive architecture, constructive dialogue models.



## 1 Introduction

The robot agent's communication capability can be regarded as one of the fundamental enablements in cognitive robotics. Given the need for collaboration and coordination of actions with the other partners as well as the agent's self-motivated exploration of the environment, it is necessary to be able to communicate one's intentions, beliefs, and desires, and for this, language is the most natural means due to its rich expressive capabilities. In HRI, action possibilities for a human are determined by the dialogue design and the models for processing inputs and generating responses, as well as by the natural language capability which allows an intuitive way to interact with the robot agent. In fact, considering the general notion of *affordance*, the robot's language capability can be said to afford intuitive interaction which is considered more usable than simple command-based protocols [8].

Affordance was originally introduced by [6] to explain human visual capability to recognize objects, and it was transferred to interface design by [19], and finally, to human-computer/robot interaction by [8], to describe the capability of a (computer) interface to readily suggest the appropriate way of behaviour. Affordance has also been used for robot architectures [15][17] to model action possibilities for a user who wishes to interact with a robot.

In the case of social robots, the use of natural language takes the interaction to a qualitatively different level and supports the robot's autonomous agent-like behaviour with dialogue features such as turn-taking, feedback, and creation of common ground. Natural dialogue interface is thus a more complex and technologically demanding design task than simply adding speech modality to the interface, and it presupposes a different frameset for the human user. In general, users tend to assign anthropomorphic features to inanimate objects like personal computers even though the objects are basically considered tools with no natural interaction capability [23]. Social robots, however, have a dual character as a tool and as an interacting agent [9], so the human-robot interaction starts to resemble human-human communicative situations.

As argued in [8], speech creates expectations for the system's ability to conduct natural language communication, and humanoid robots reinforce such expectations with their human-like appearance, including aspects like personality [21] and even stereotypical roles and gender [25]. The need for natural language interaction and affordable interfaces thus involves dialogue modelling concerning language analysis and interpretation, and the robot agent should also be able to understand multimodal sensory information that it receives from its environment. Conversely, it should be able to produce behaviour that matches requirements of a relevant and coherent response, combining spoken language and multimodality (gestures, gaze, body posture). An important aspect of this work is to design a dialogue architecture which supports natural interaction and allows experimentation with various multimodal modules so as to explore human experience with humanoid robots and address larger societal needs to find new ways to improve the robot agents' acceptance and usability in society.

In this paper, the discussion focuses on the architecture that supports these requirements. Section 2 briefly describes the Constructive Dialogue Model and the Standard Model of Cognitive Architecture. Section 3 shows the intended integration of the models, and Section 4 provides short discussion of the topics and future work.

## 2 Architectures for Cognitive Robots

### 2.1 Constructive Dialogue model

The Constructive Dialogue Model (CDM) [8] is a complementary architecture to cognitive architectures (ACT-R, Soar, Standard Model [13]) which do not explicitly concern dialogue communication. CDM can be implemented on top of the cognitive perception-action modules as a component responsible for the higher-level reasoning on verbal and multimodal communication. It is chosen because of its focus on natural language dialogues and because of its links to cognitive aspects of interaction (communicative enablements). Also, it has been used in robot applications [12][27].

CDM is a conceptual and operational framework which regards conversational interactions as cooperative activities through which the interlocutors build common ground (cf. similar approaches in [4][20][26]). In CDM, the participants are regarded as rational agents, engaged in cooperative activity within which they aim to achieve their communicative goals using dialogue acts which convey information about their intentions and task topics. The agents exchange new information on the relevant top-

ics in order to construct mutual understanding and coordinate their actions (see [10] for dialogue management issues in general).

Figure 1 shows how conversations progress in a cyclic manner as the participants produce utterances and check various enablements for communication [1] to maintain interaction and monitor its progress. For instance, the agents must be in contact and aware of the partner's attempt to communicate, by paying attention to (multimodal) signals that indicate their willingness to interact [5]. The agents must also perceive the emitted vocal and visual signals as communicative signals, i.e. recognize them having been produced with an intention to convey meaning. The agents must also intend to engage themselves in the communication, i.e. make an effort to understand the partner's message and intentions, and to produce their own reaction. Reaction encodes new information about the agent's current viewpoint in verbal or physical actions. It changes the current state of the world and requires the agents to restart their reasoning with the new situation. The cycle continues until the conversation is finished by the agents mutually agreeing to stop, or for another reason.

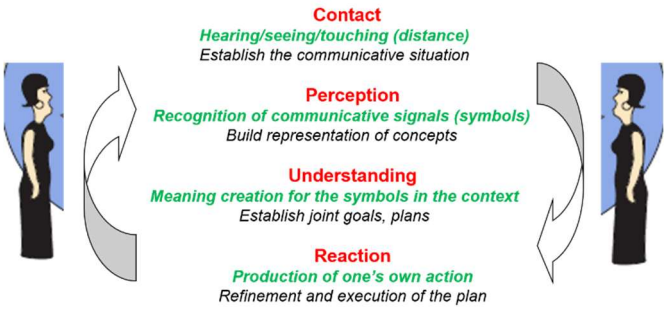


*Figure 1 Enablements for Constructive Dialogue Model [8].*

Rationality refers to the agent's ability to make decisions and deliberate on situationally appropriate actions (in AI, such agents have been called BDI agents), and it also considers the agent's affective state which influences the agent's reasoning. Emotions [3] are not explicitly represented in the architecture, since they are assumed to be manifestations of the agent's internal state: the levels of arousal and valence of the agent's affects are inherent to the agent's general activity rather than computed by a particular emotion component. In fact, emotional activity can be regarded as one of the connection points of CDM to the cognitive architectures under the assumption that the processing of input signals results in an internal state which determines the emotional quality of the agent's response.

## 2.2 Cognitive Robotics Models

Two cognitive models are shown in Figure 2: ACT-R [2] and the Standard Model [13] (we do not discuss the third main framework, SOAR here). The focus is to model human behaviour based on the perception of the environment and the (motor) actions that the agent can take as the result of its reasoning. Consequently, studies have dealt with the visual and auditive systems and their functionality in the context of short-term (working) and long-term (declarative) memory. Important research has also fo-

cused on linguistic resources and knowledge representation for reasoning and conceptual categorization tasks. For instance, the integrated knowledge representation system Dual-PECCS [14] uses two different sorts of common-sense reasoning, prototypical and exemplars-based categorization, to allow knowledge acquisition and development of Conceptual Space representations for a variety of tasks.

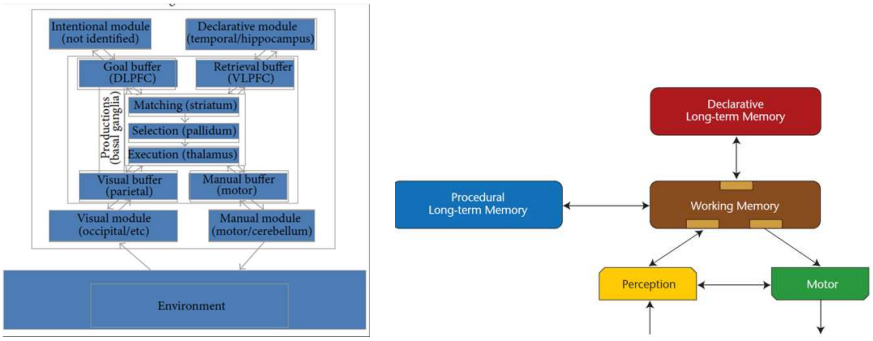


*Figure 2 ACT-R Architecture (left) and the Standard Model (right), from [9].*

Knowledge representation is an important aspect of cognitive processing and has been a topic of much debate (centralized or distributed processing, symbolic or connectionist representation, procedural or declarative knowledge, etc.). From the dialogue point of view, there is a need for uniform representation of the meaning in order to allow higher-level modules to operate on meaningful chunks of the incoming information. While the overall view of the cognitive models involves two memory components (Working Memory and procedural/declarative Long-term Memory), it is likely that some kind of hybrid knowledge representation is needed to toggle between declarative and procedural knowledge in the system's working memory. Moreover, since the agent needs to *ground* its knowledge in the physical world [7], an interim representation seems necessary to connect the concepts stored in the agent's memory to its dynamic perception of the world. In fact, in recent years, the *ProxyType Theory* [22] has been proposed to cater for heterogeneity in concept representations and to address issues concerning the interaction between Long-term and Working Memory. According to the theory, the process of proxification manages conceptual structures into temporary constructs in Working Memory using heterogeneous representations: activation of a concept category in Long-term Memory (which contains networks of representations) results in the concept's activation as token representation in Working Memory (as a "proxy" for the concept). Concept categories are complex networks of (neural) activations among the network elements, and they are constructed over time via perceptual interactions of the agent with the environment which results in repeated activations of relevant elements in the connected networks. (The network elements are causally connected since activation of an element will cause the activation of the connected elements in Long-term Memory, and the tokening of the concept category in Working Memory.)

The ultimate goal for a robot agent is to learn via experience and be able to adapt to the dynamically changing world. The agent's continuous learning of new concepts

and skills is an important part of interaction management and allows the agent to coordinate action in order to adjust to its environment. However, the effectiveness of interactive learning depends on the quality of the interactions. So far most frequently used settings have included designer-controlled ways of interaction which are based on linear learning and scripted interaction sequences. Free natural dialogue interactions provide new challenges by leveraging deep interactive learning and building of competences through interactions: besides the technically demanding aspects related to recognition and processing of various multimodal signals, such interactions presuppose understanding of the partner's intentions and partly developed skills, as well as social aspects of interaction. They require specific models for interaction through which social behaviours are learned and saved as persistently growing experience of the world. The goal also incorporates the issues studied in Theory of Mind [28] to construct a shared context for mutual understanding and shared context, which have also been some of the main issues in cooperative dialogue approaches.

## 3 Integrated Architecture

While dialogue modelling also subscribes to the goals related to knowledge representation and learning, the main focus is on the models of interaction management. Given the cascaded model of communicative enablements as presented above with Contact, Perception, Understanding, and Reaction, it is easy to see how to extend the Standard Cognitive Architecture by the CDM dialogue component which deals with the processing and management of interaction. The integrated architecture in Figure 3 shows the basic requirements for spoken dialogue models and integration points with CDM.

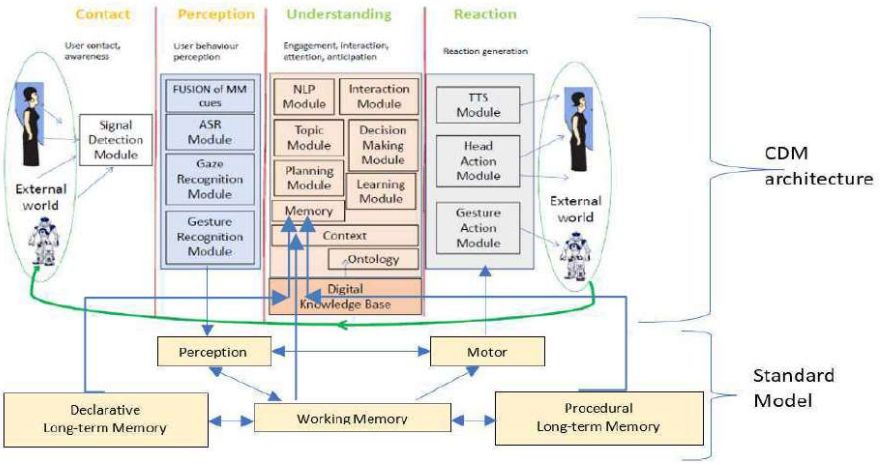


*Figure 3 Cognitive Robot Architecture which integrates the CDM dialogue model and the Standard Model of cognitive processing (from [9]).*

The integration points for the Perception and Motor control components of the Standard Architecture are the CDM modules related to the enablements of Perception and Reaction, whereas Long-Term Memory (both declarative and procedural) and Work-

ing Memory correlate with the modules in CDM Understanding. The detailed CDM Understanding modules encode the system's procedural knowledge for the analysis of the user's behaviour and deciding what to do next, as well as modules for three types of knowledge: the CDM Knowledge Base stores the robot's (long-term) knowledge of the domain, of the user, and of the world in general, while the CDM Memory stores the system's knowledge of the past dialogue events it has been engaged in, and the CDM Context models the immediate dialogue context and the (short-term) state of attention of the agent. It is worth noticing that the architecture makes an explicit distinction between semantic and episodic memory: semantic memory is scattered among the system components (e.g. language grammar belongs to the NLP Module and planning rules to the Planning Modules), whereas the CDM Memory is episodic and refers to a designated part of the agent's knowledge where the previous sessions with a particular user are saved, and from where chunks of knowledge are retrieved for dialogue processing if the partner is identified as a returning user to interact with. All knowledge sources are connected to the other processing modules via Ontology, which provides semantic links between the Knowledge Base entities and linguistic concepts. As in Dual-PECCS [14], it is also possible to use other linguistic resources to provide an interface between the linguistic and the conceptual knowledge.

The double nature of the robot as an agent and as a computer system sets requirements for the dialogue model. As an agent, the robot is perceived as a communicating partner, and as a computer system, it has access to vast digital information which it can also share with other agents through its connection to Internet (IoT [24]). The integrated architecture described above does not specifically deal with interactions in the ubiquitous environment, but it is possible to include sensor information as input through specific perception devices, and then visualise the data and process it as normal (cf. [29]). However, the ubiquitous environment can also drastically change the knowledge available for the agent, e.g. digital database is modified according to new data, and in these cases, the robot agent needs to possess procedural knowledge of how to cope with unexpected, unspecified, or underspecified situations. This paper does not discuss these issues but emphasises that probabilistic modelling of knowledge together with the agent's capability to learn are crucial in their realisation.

## 4 Discussion and Future Work

In the field of Human-Robot Interaction (HRI), one of the important and much discussed topics is the notion of Uncanny Valley [18], whereby the robot's human-like appearance is correlated with the acceptance of the robot as an interactive partner. At one end, we have robot agents which look and behave like human agents, while at the other end, the interaction partner is clearly a non-human agent which may exhibit different levels of human-likeness. The hypothesis states that the acceptance of agent applications increases when going from less human-like agents towards close to human-level behaviour, but there is a sudden drop in the acceptance when the robot agent reaches almost the same level of behaviour as the human. The Uncanny Valley phenomenon has since been shown to appear as a result of a mismatch in cognitive

categorization [16] of what is considered similar to but not exactly the same as the prototypical conversing agent (namely the human). Contradiction between typical members of a class and entities which deviate from them usually causes uncomfortableness, fear and resistance, and explains why talking robots create similar reactions. In HRI, such cognitive mismatches are commonly triggered by the robot's appearance and look, but also by its capability to interact with humans.

The proposed architecture is considered a valuable first step to achieve natural dialogue interactions in HRI, and make the robot behave in a more natural manner. There are several aspects that can be further specified to experimentally validate the architecture and make the contributions visible, especially with respect to the integration of cognitive architectures into the CDM dialogue model. On the theoretical side, appropriate knowledge representation and integration of multimodal input (gestures, eye-gaze) will be elaborated, and the ProxyType theory of concept representation will be investigated further. On the practical robotics side, the questions of how to develop socially competent robots and use novel AI technology to alleviate problems in the modern society will be explored.

Future work will proceed using a top-down and bottom-up (TDBU) methodology: this aims to combine the theoretical model of interaction (top-down) with the automatic recognition techniques and data analysis (bottom-up). The TDBU methodology uses novel technology to provide an objective basis for detecting and segmenting elements in the interaction flow, while the theoretical views of human observations and annotations are used for the interpretation and parameter setting. Speech recognizers, parsers, eye-trackers, movement detectors, etc. are used to segment signals and provide bottom-up knowledge to trace gaze, face, and body, while the theoretical view of dialogue modelling and communicatively important signals are used to explore meaningful correlations and regularities in the (big) data. Deep learning techniques and statistical correlations are used to develop such models.

To explore how social robots can assist humans in various every-day tasks, the work will continue in an interdisciplinary manner using experimental methods from cognitive and social sciences to study user experience and engagement in social human-robot situations. Experimental design can consist of different types of robot agents ("personalities") and of strategic profiling with respect to such issues as active narration vs. passive guidance, use of gestures, amount of feedback, etc. to compare the user's experience and engagement with the robot agent in the selected scenarios.

## Acknowledgements

The author wishes to thank colleagues for discussions on ontologies and dialogue modelling, and Antonio Lieto for discussions related to Dual-PECCS. The study is based on the results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

## References


1. Allwood, J. (1976). *Linguistic Communication as Action and Coordination*. Gothenburg Monographs in Linguistics, 2, Göteborg.
2. Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, 111(4):1036.
3. Barret, L.F., Lewis, M., Haviland-Jones, J.M. (2008, Eds.) *Handbook of Emotions*. Guildford Press, New York.
4. Clark, H. H., Schaefer, E. F. (1987). Collaborating on contributions to conversation. *Language and Cognitive Processes*, 2, pp. 19-41.
5. Feldman R., Rim B. (1991). *Fundamentals of Nonverbal Behavior*. Cambridge Univ. Press
6. Gibson, J.J. (1979). *The Ecological Approach to Visual Perception*. Houghton Mifflin.
7. Harnad, S. (1990). The symbol grounding problem. *Physica D* 42: 335–346.
8. Jokinen, K. (2009). *Constructive Dialogue Modelling – Speech Interaction with Rational Agents.* John Wiley, Chichester, UK.
9. Jokinen, K. (2018). Dialogue Models for Socially Intelligent Robots. The 10th International Conference on Social Robots, Qingdao, China.
10. Jokinen, K., McTear, M. (2009). *Spoken Dialogue Systems*. Morgan and Claypool.
11. Jokinen, K., Nishimura, S., Watanabe, K., Nishimura, T.(2018). Human-Robot Dialogues for Explaining Activities. Proceedings of IWSDS-2018, Singapore.
12. Jokinen. K., Wilcock, G. (2013). Multimodal Open-Domain Conversations with the Nao Robot. In: Natural Interaction with Robots, Knowbots and Smartphones: Putting Spoken Dialogue Systems into Practice, pages 213–224. Springer, New York.
13. Laird, J.E. Lebiere, C., Rosenbloom, P.S. (2017). A Standard Model of the Mind: Toward a Common Computational Framework Across Artificial Intelligence, Cognitive Science, Neurosciene, and Robotics. *AI Magazine* 38(4):13-26.
14. Lieto, A., Radicioni, D.P., Rho, V. (2016). Dual PECCS: a cognitive system for conceptual representation and categorization. *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 1–20.
15. Marin-Urias, L.F., Sisbot, E.A., Pandey, A.K., Tadakuma, R., Alami, R. (2009). Towards shared attention through geometric reasoning for human robot interaction. In: Humanoids 2009. The 9th IEEE-RAS International Conference on Humanoid Robots, pp. 331-336.
16. Moore, R. (2014). A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3499759/
17. Moratz, R., Tenbrink, T. (2008). Affordance-Based Human-Robot Interaction. Towards Affordance-Based Robot Control. Lecture Notes in Computer Science 4760 pp. 63 – 76.
18. Mori M. (1970). The Uncanny Valley. *Energy*, 7(4), 33–35.
19. Norman, D.A. (1988). *The Psychology of Everyday Things*. Basic Books: New York.
20. Nooraei, B., Rich C., Sidner, C. (2014). A Real-Time Architecture for Embodied Conversational Agents: Beyond Turn-Taking. The 7th Int. Conf. on Advances in Computer-Human Interactions.
21. Okada, S., Nguyen, L.S., Aran, O., Gatica-Perez, D. (2019). Modeling Dyadic and Group Impressions with Intermodal and Interpersonal Features. *ACM Transactions on Multimedia Computing Communication Applications* 15, 1s, Article 13 (January 2019).
22. Prinz, J.J. (2002). *Furnishing the mind: Concepts and their perceptual basis*. MIT Press.
23. Reeves, N., Nass C. (1996). *The Media Equation: How people treat computers, television, and new media like real people and places*. New York: Cambridge University Press.
24. Smith, I. G. (Ed. 2012). The Internet of Things 2012: New Horizons. IERC-Internet of Things European Research Cluster. Halifax, U.K.

25. Tay, B., Jung, Y., Park, T. (2014). When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction. *Computers in Human Behaviour*, Vol 38, pp. 75-84.
26. Traum, D.R., Allen, J.F. (1994). Discourse obligations in dialogue processing. Proceedings of the 32nd Annual Meeting of ACL, pp. 1–8. Morristown, NJ, USA.
27. Wilcock, G., Jokinen, K. (2015). Multilingual WikiTalk: Wikipedia-based talking robots that switch languages. Proceedings of the SIGDIAL 2015 Conference, pp. 162-164.
28. Wimmer, H., Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* 13:103–128.
29. Yoshida, Y., Nishimura, T., Jokinen, K. (2018). Biomechanics for understanding movements in daily activities. Procs of the LREC Workshop "Language and Body in Real Life".