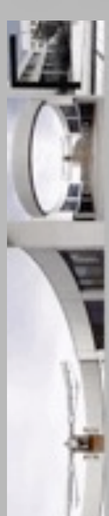


Talking Robots

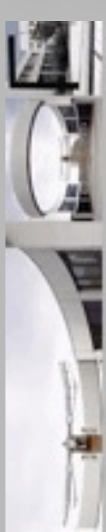
Situated Dialogue Processing for Human-Robot Interaction

Dr.ir. Geert-Jan M. Kruijff

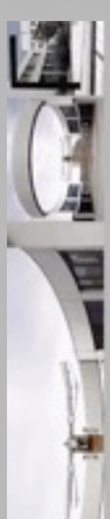
Talking Robots
@the Language Technology Lab
DFKI GmbH, Saarbrücken

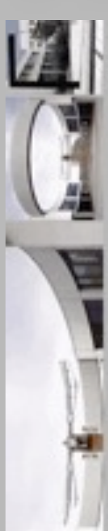
http://talkingrobots.dfki.de

Robots

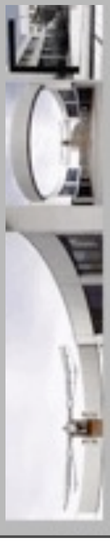


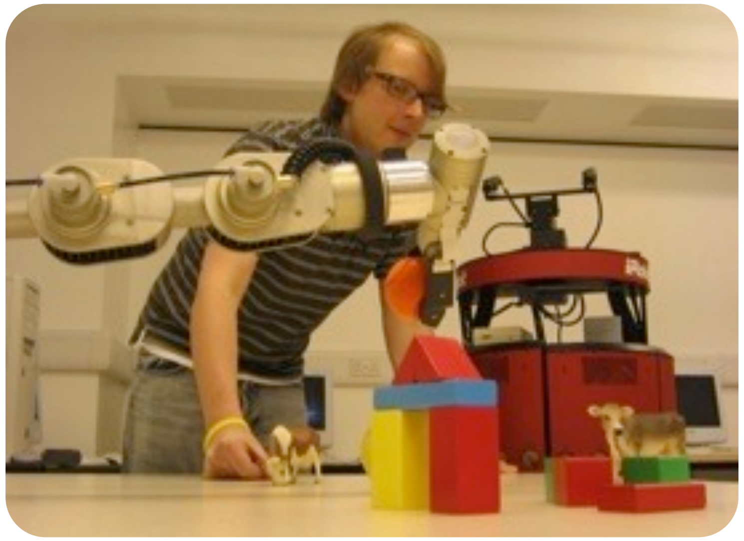
TALKING ROBOTS
LANGUAGE TECHNOLOGY, DFKI



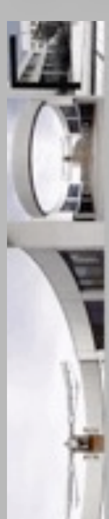


ASSIST PEOPLE



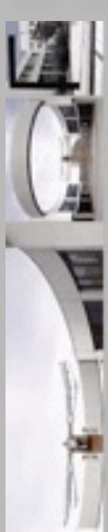



Kellogg's
CORN FLAKES



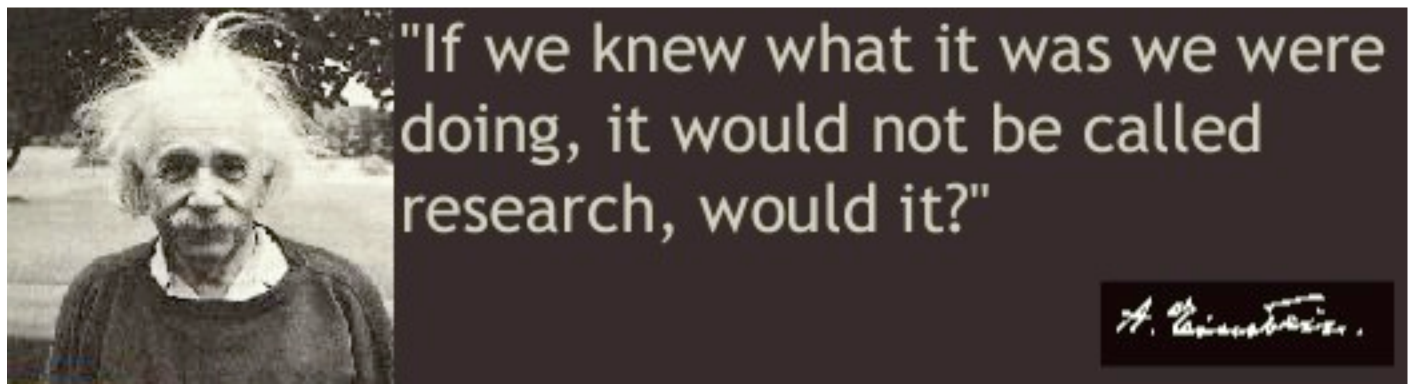
CHALLENGE

$\nRightarrow$

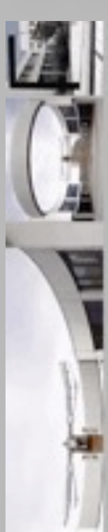


Bridge – Understand – Help

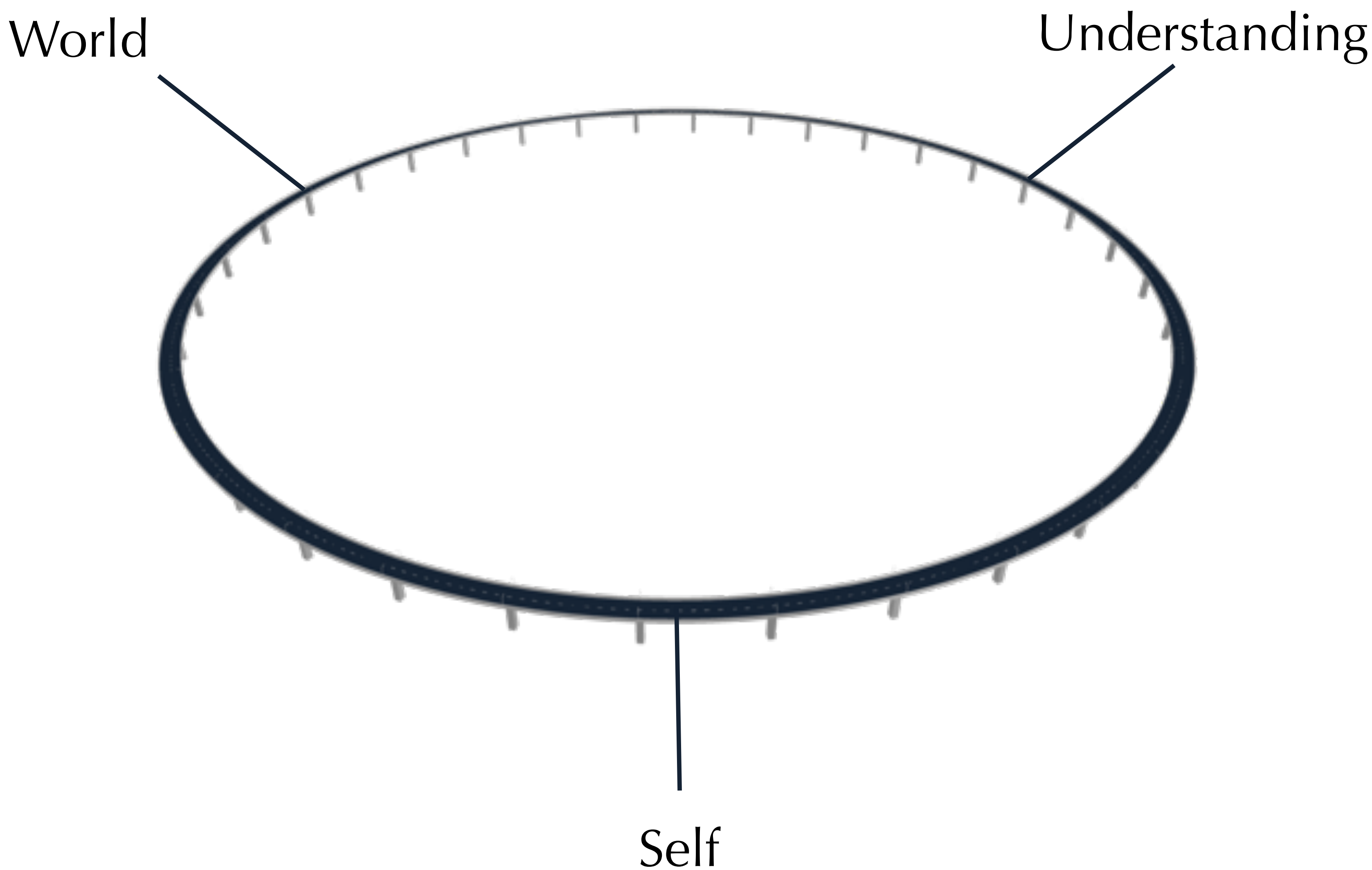

I THINK YOU SHOULD BE MORE SPECIFIC HERE IN STEP TWO
THEN A MIRACLE OCCURS ...
S. Harris

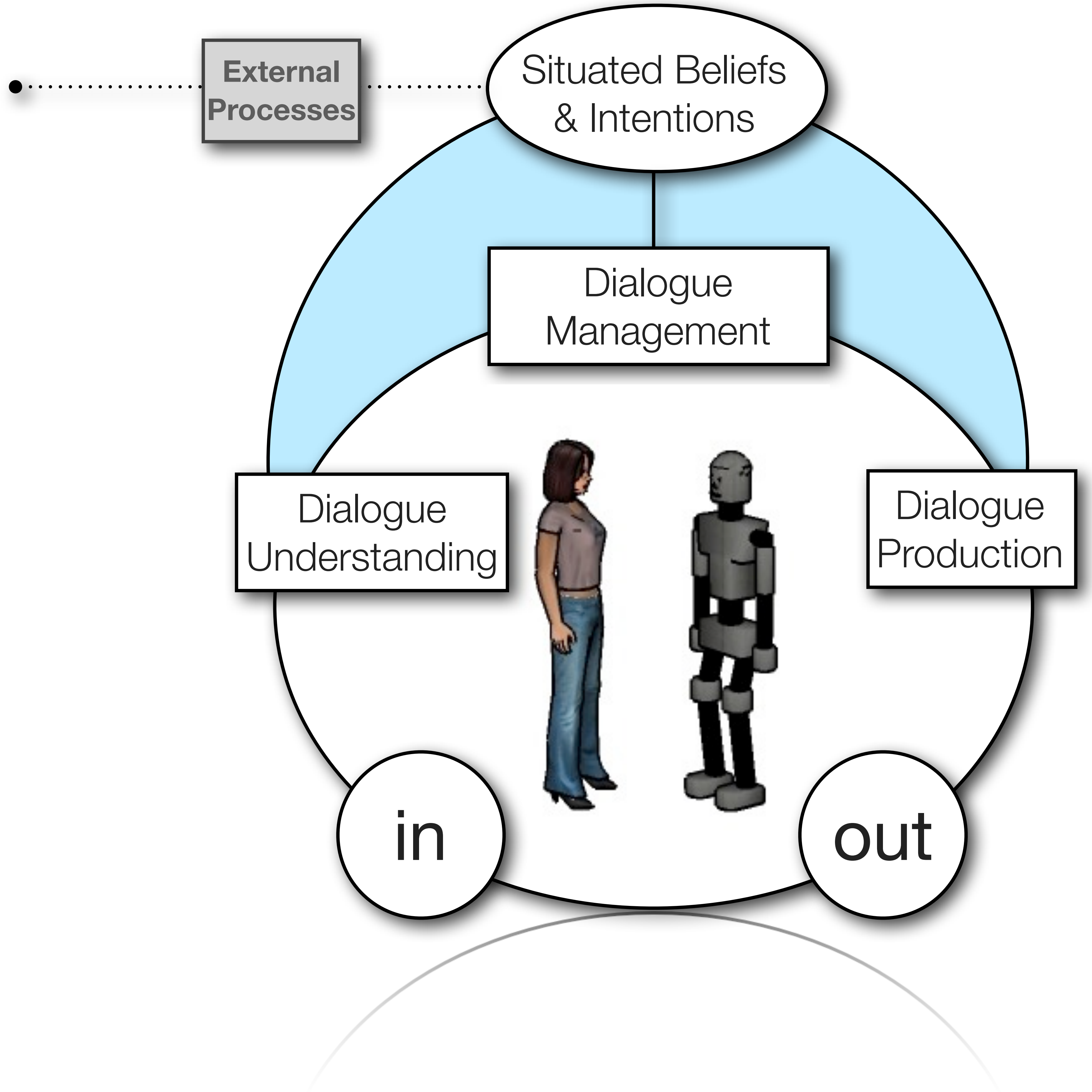

"If we knew what it was we were doing, it would not be called research, would it?"

Talking Robots
Language Technology, DFKI



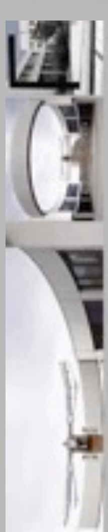
World – Other – Self


World
Understanding
Self


External Processes
Situated Beliefs & Intentions
Dialogue Management
Dialogue Understanding
Dialogue Production
in
out

Bridge = Mediation

- Ontology-based mediation



Bridge = Mediation

- Ontology-based mediation

Information Fusion For Visual Reference Resolution In Dynamic Situated Dialogue

Geert-Jan M. Kruijff[1], John D. Kelleher[2], and Nick Hawes[3]

[1] Language Technology Lab, DFKI GmbH
gj@dfki.de,
WWW home page: http://www.dfki.de/~gj
[2] Dublin Institute of Technology
John.Kelleher@comp.dit.ie,
WWW home page: www.computing.dcu.ie/jkelleher/
[3] School of Computer Science, University of Birmingham
n.a.hawes@cs.bham.ac.uk,
WWW home page: http://www.cs.bham.ac.uk/~nah

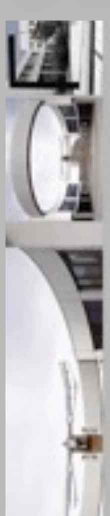
Abstract. Human-Robot Interaction (HRI) invariably involves dialogue about objects in the environment in which the agents are situated. The paper focuses on the issue of resolving discourse references to such visual objects. The paper addresses the problem using strategies for *intra-modal fusion* (identifying that different occurrences concern the same object), and *inter-modal fusion*, (relating object references across different modalities). Core to these strategies are sensori-motoric coordination, and ontology-based mediation between content in different modalities. The approach has been fully implemented, and is illustrated with several working examples.

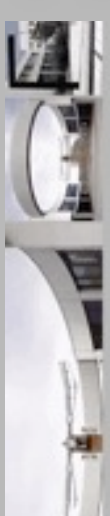
1 Introduction

The context of this work is the development of dialog systems for human-robot collaboration. The framework presented in this paper addresses a particular aspect of situated dialog, namely reference resolution. Reference resolution in situated dialog is a particular instance of the anchoring problem [Coradeschi and Saffiotti, 2003]: how can an artificial system create and maintain correspondences between the symbols and sensor data that refer to the same physical object?

In a dialog, human participants expect their partner to construct and maintain a model of the evolving linguistic context. Each referring expression used in the dialog introduces a representation into the semantics of its utterance. This representation must be bound to an element in the context model in order for the utterance's semantics to be fully resolved. Referring expressions that access a representation in the context are called *anaphoric*. In a *situated* dialog, human participants expect their partner to not only construct and maintain a model of the linguistic discourse, but also to have full perceptual knowledge of the environment. This introduces a form of reference, called *exophoric* reference. Exophoric references denote objects that have entered the dialog context through a non-linguistic modality (such as vision), but have not been previously evoked into the context. Consequently, for a robot to participate in a situated dialog,

Talking Robots
Language Technology, DFKI





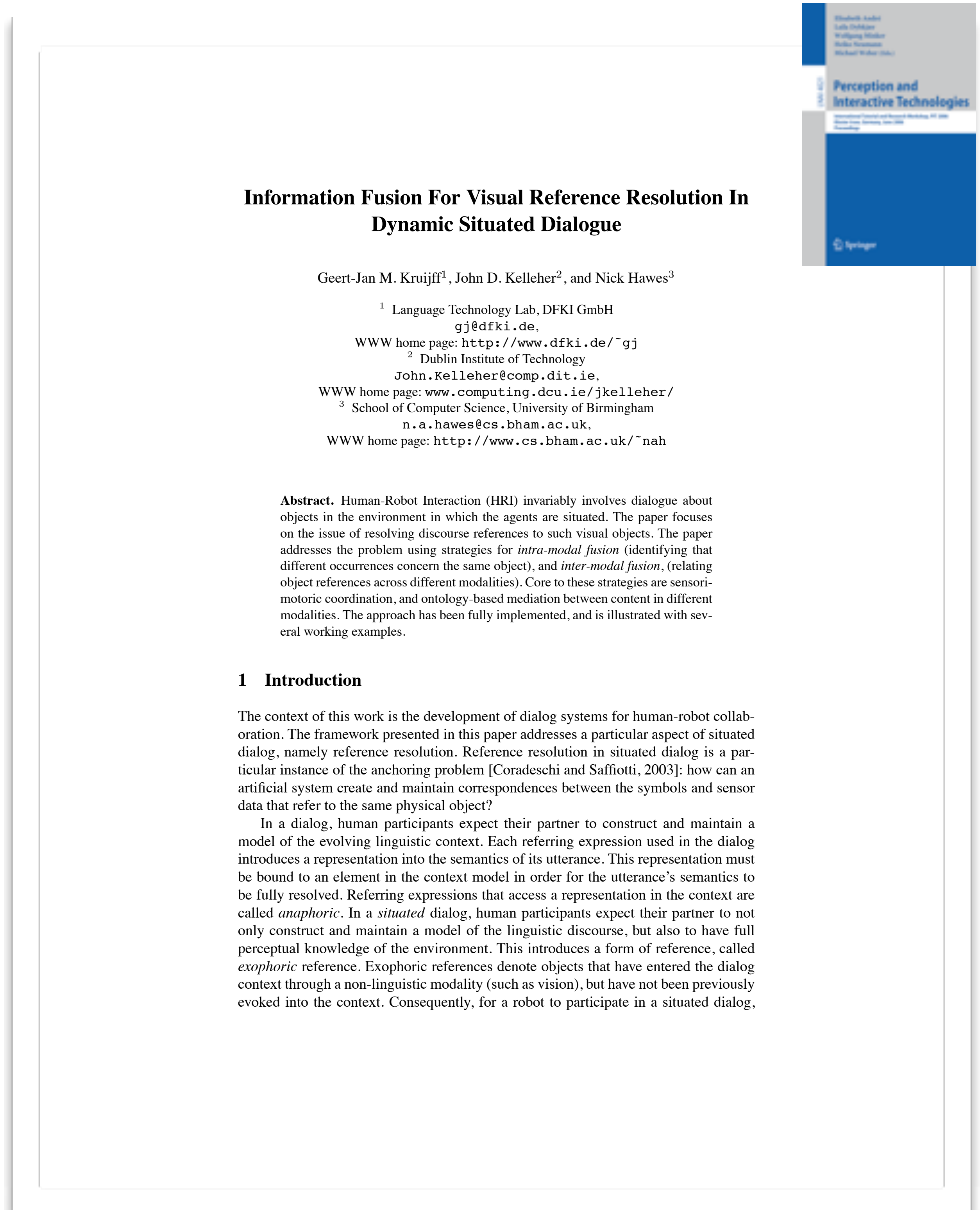
Bridge = Mediation

- Ontology-based mediation

Conceptual meaning:
concept(**box**) & *instance*(l1,**box**) & i1 $\rightleftarrows$ l1

Mapping

Linguistic meaning:
@{i1:object}(**box**)



Information Fusion For Visual Reference Resolution In Dynamic Situated Dialogue

Geert-Jan M. Kruijff[1], John D. Kelleher[2], and Nick Hawes[3]

[1] Language Technology Lab, DFKI GmbH
gj@dfki.de,
WWW home page: http://www.dfki.de/~gj
[2] Dublin Institute of Technology
John.Kelleher@comp.dit.ie,
WWW home page: www.computing.dcu.ie/jkelleher/
[3] School of Computer Science, University of Birmingham
n.a.hawes@cs.bham.ac.uk,
WWW home page: http://www.cs.bham.ac.uk/~nah

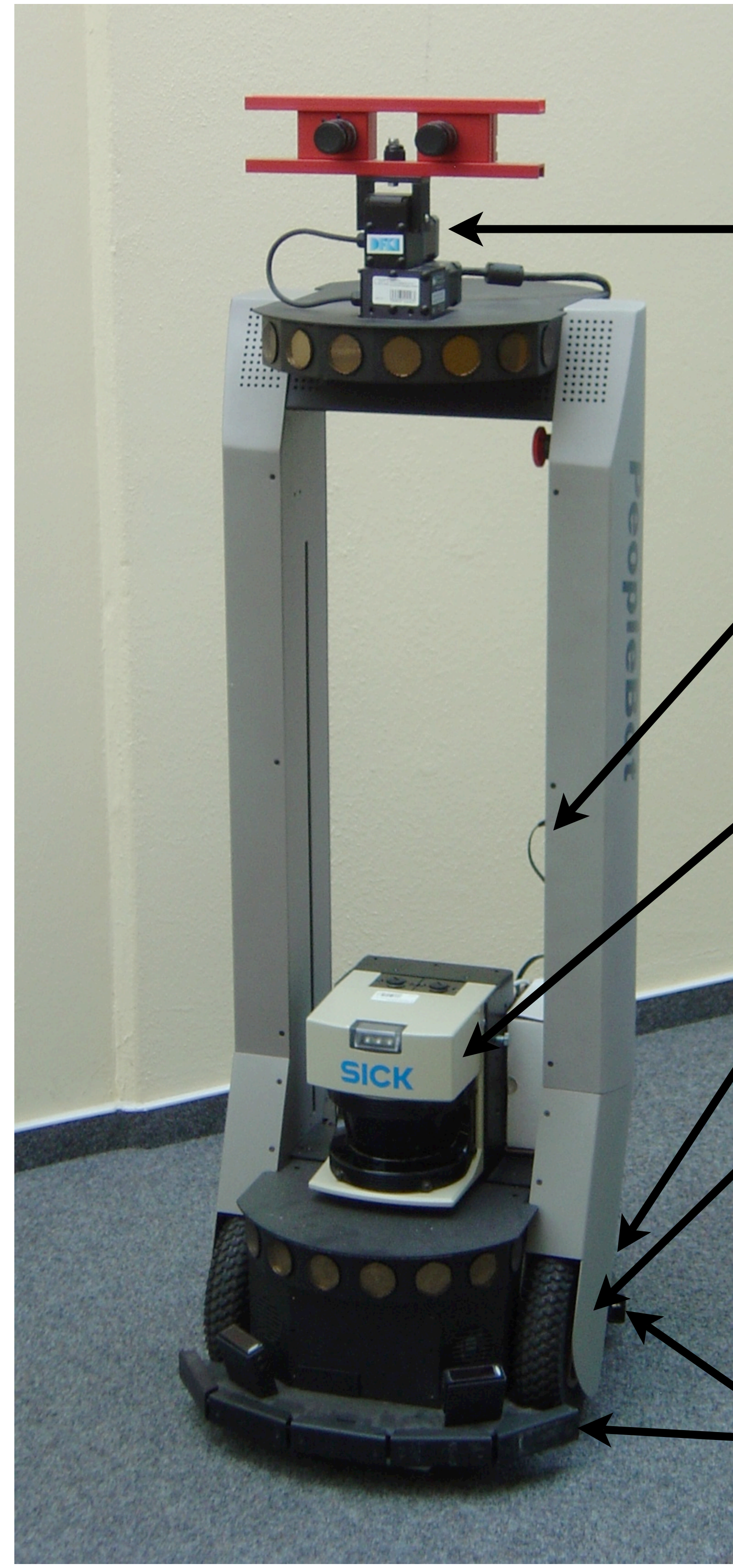
Abstract. Human-Robot Interaction (HRI) invariably involves dialogue about objects in the environment in which the agents are situated. The paper focuses on the issue of resolving discourse references to such visual objects. The paper addresses the problem using strategies for *intra-modal fusion* (identifying that different occurrences concern the same object), and *inter-modal fusion*, (relating object references across different modalities). Core to these strategies are sensori-motoric coordination, and ontology-based mediation between content in different modalities. The approach has been fully implemented, and is illustrated with several working examples.

1 Introduction

The context of this work is the development of dialog systems for human-robot collaboration. The framework presented in this paper addresses a particular aspect of situated dialog, namely reference resolution. Reference resolution in situated dialog is a particular instance of the anchoring problem [Coradeschi and Saffiotti, 2003]: how can an artificial system create and maintain correspondences between the symbols and sensor data that refer to the same physical object?

In a dialog, human participants expect their partner to construct and maintain a model of the evolving linguistic context. Each referring expression used in the dialog introduces a representation into the semantics of its utterance. This representation must be bound to an element in the context model in order for the utterance's semantics to be fully resolved. Referring expressions that access a representation in the context are called *anaphoric*. In a *situated* dialog, human participants expect their partner to not only construct and maintain a model of the linguistic discourse, but also to have full perceptual knowledge of the environment. This introduces a form of reference, called *exophoric* reference. Exophoric references denote objects that have entered the dialog context through a non-linguistic modality (such as vision), but have not been previously evoked into the context. Consequently, for a robot to participate in a situated dialog,

Example: Mapping



Example: Mapping

Situated Dialogue and Spatial Organization: What, Where... and Why?

Geert-Jan M. Kruijff[1]; Hendrik Zender[1]; Patric Jensfelt[2] & Henrik I. Christensen[2]
[1]Language Technology Lab, German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany
[2]Centre for Autonomous Systems, Royal Institute of Technology (KTH), Stockholm, Sweden
gj@dfki.de

Abstract: *The paper presents an HRI architecture for human-augmented mapping, which has been implemented and tested on an autonomous mobile robotic platform. Through interaction with a human, the robot can augment its autonomously acquired metric map with qualitative information about locations and objects in the environment. The system implements various interaction strategies observed in independently performed Wizard-of-Oz studies. The paper discusses an ontology-based approach to multi-layered conceptual spatial mapping that provides a common ground for human-robot dialogue. This is achieved by combining acquired knowledge with innate conceptual commonsense knowledge in order to infer new knowledge. The architecture bridges the gap between the rich semantic representations of the meaning expressed by verbal utterances on the one hand and the robot's internal sensor-based world representation on the other. It is thus possible to establish reference to spatial areas in a situated dialogue between a human and a robot about their environment. The resulting conceptual descriptions represent qualitative knowledge about locations in the environment that can serve as a basis for achieving a notion of situational awareness.*

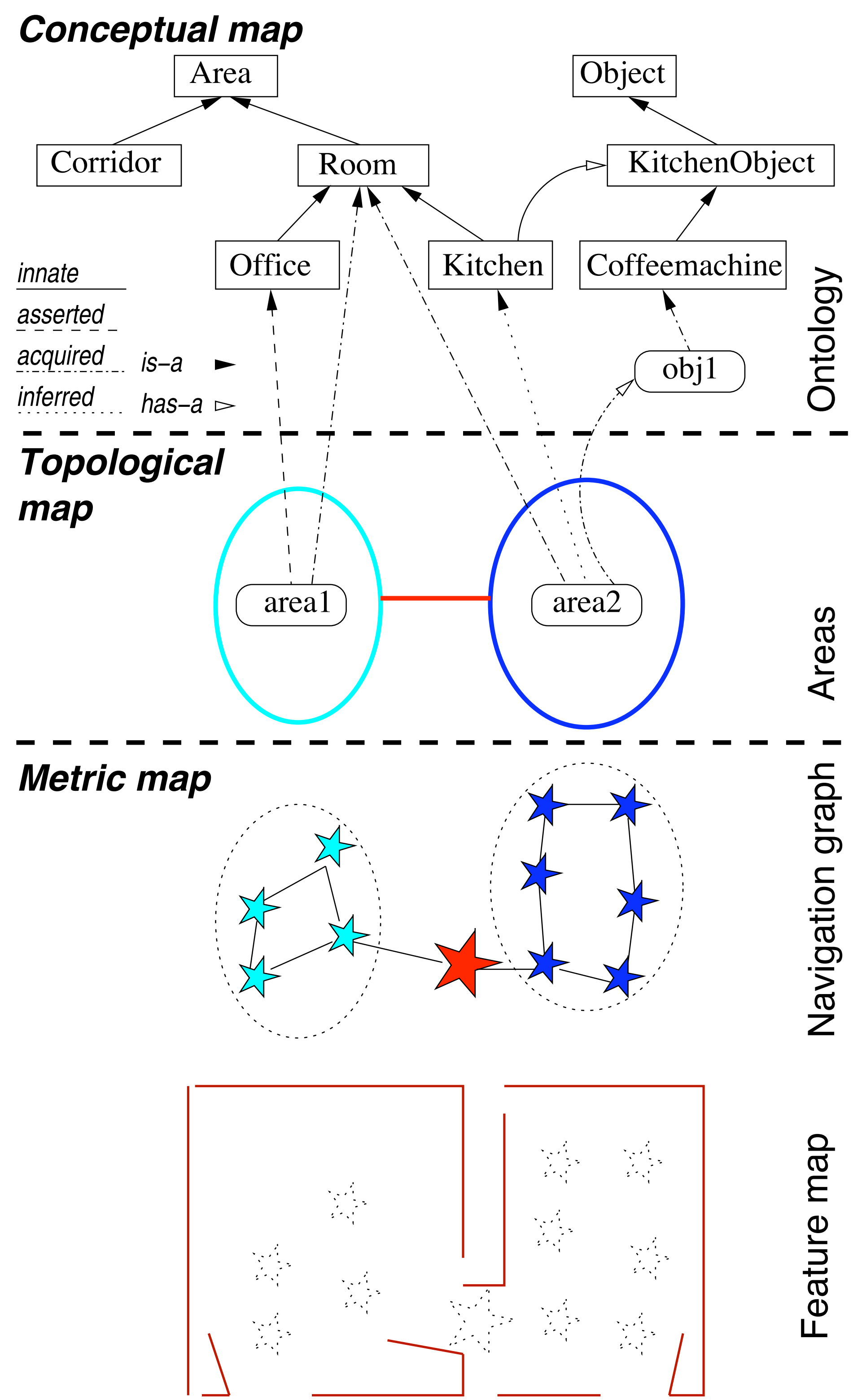
Keywords: *Human-Robot Interaction, Conceptual Spatial Mapping, Situated Dialogue*

1. Introduction

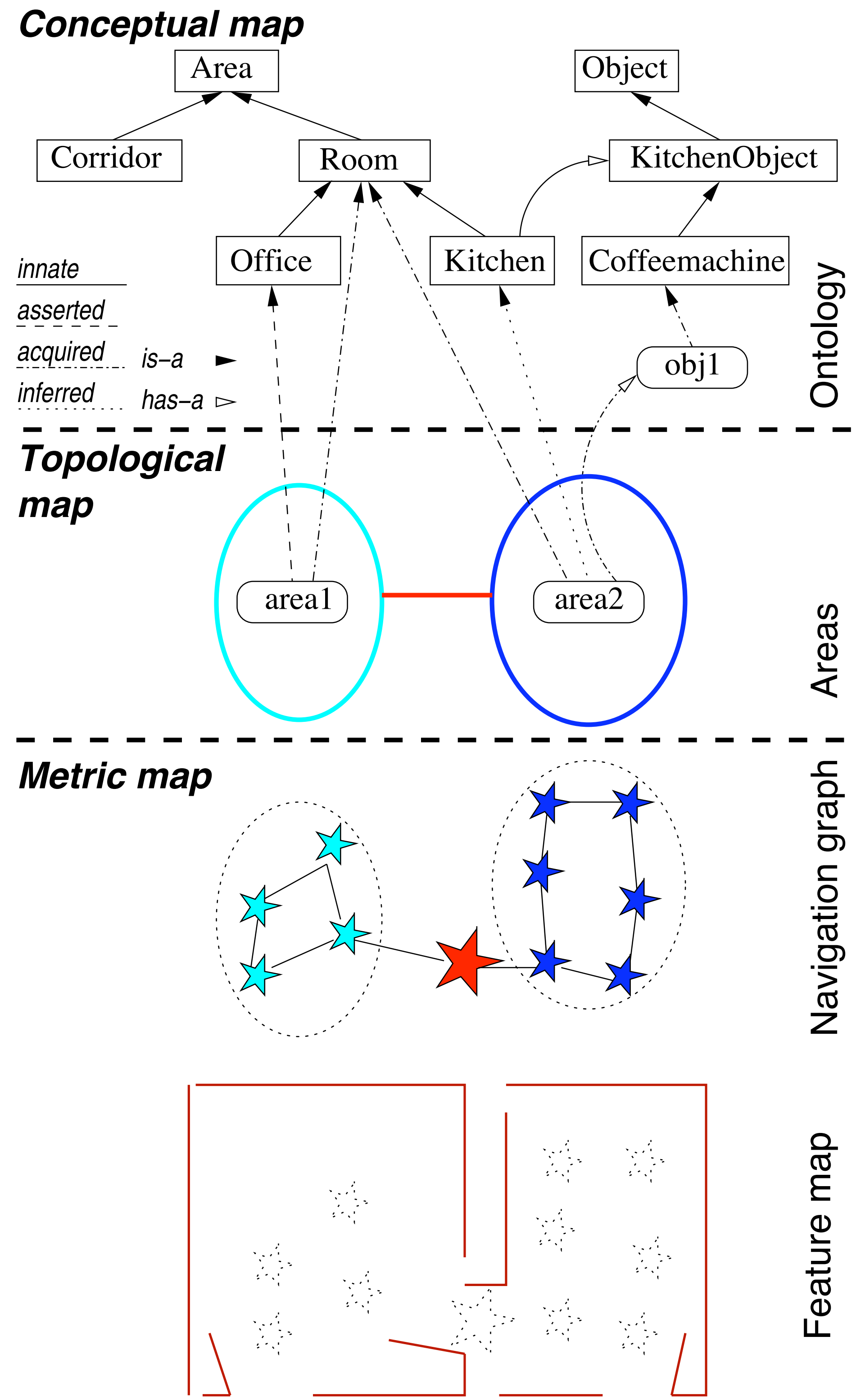
More and more robots find their way into environments where their primary purpose is to interact with humans to help and solve a variety of service-oriented tasks. Particularly if such a service robot is mobile, it needs to have an understanding of the spatial and functional properties of the environment in which it operates. The problem we address is how a robot can acquire an understanding of the environment so that it can autonomously operate in it, and communicate about it with a human. We present an architecture that provides the robot with this ability through a combination of human-robot interaction and autonomous mapping techniques. It captures various functions that independently performed Wizard-of-Oz studies have observed to be necessary for such a system. Several case studies have been conducted to test and evaluate the resulting integrated system.
The main issue is how to establish a correspondence between how a human perceives spatial and functional aspects of an environment, and what the robot autonomously learns as a map. Most existing approaches to robot map building, or Simultaneous Localization And Mapping (SLAM), use a metric representation of space. Humans, though, have a more qualitative, topological perspective on spatial organization (McNamara, 1986). We adopt an approach in which we build a multi-layered representation of the environment, combining metric maps and topological graphs (as an abstraction over geometrical information), like (Kuipers, 2000). We extend these representations with conceptual descriptions that capture aspects of spatial and functional organization. The robot obtains these descriptions either through interaction with a human, or through inference combining its own observations (*I see a coffee machine*) with ontological knowledge (*Coffee machines are usually found in kitchens, so this is likely to be a kitchen!*). We store objects in the spatial representations, and so associate the functionality of a location with that of the functions of the objects present there. A core characteristic of our approach is that we analyze each utterance to obtain a representation of the meaning it expresses, and how it (syntactically) conveys that meaning – rather than just doing for example keyword spotting. This way, we can properly handle the variety of ways in which people may express assertions, questions, and commands. Furthermore, having a representation of the meaning of the utterance we can combine it with further inferences over ontologies to obtain a complete conceptual description of the location or object being talked about. This way we can ground situated dialogue in the situational awareness of the robot.
Following (Topp & Christensen, 2005) and (Topp et al., 2006), we talk about *Human-Augmented Mapping* (HAM) to indicate the active role that human-robot interaction plays in the robot's acquisition of qualitative spatial knowledge. In §2 we discuss various observations that independently performed Wizard-of-Oz studies have made on typical interactions for HAM scenarios, and we

International Journal of Advanced Robotic Systems, Vol. x, No. y (200z)
ISSN 1729-8806, pp. first-last (leave this section unchanged)

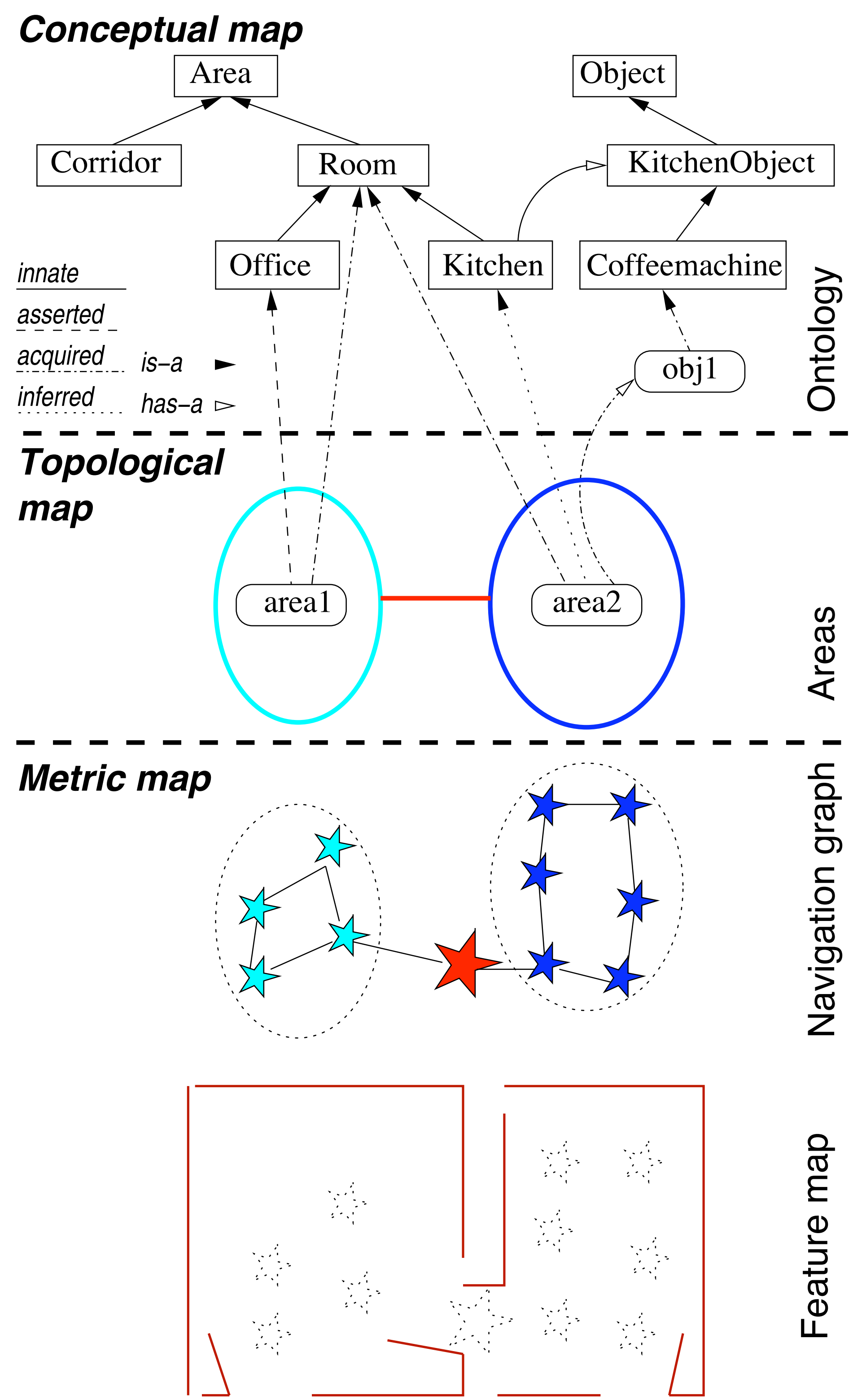
Example: Mapping



Example: Mapping

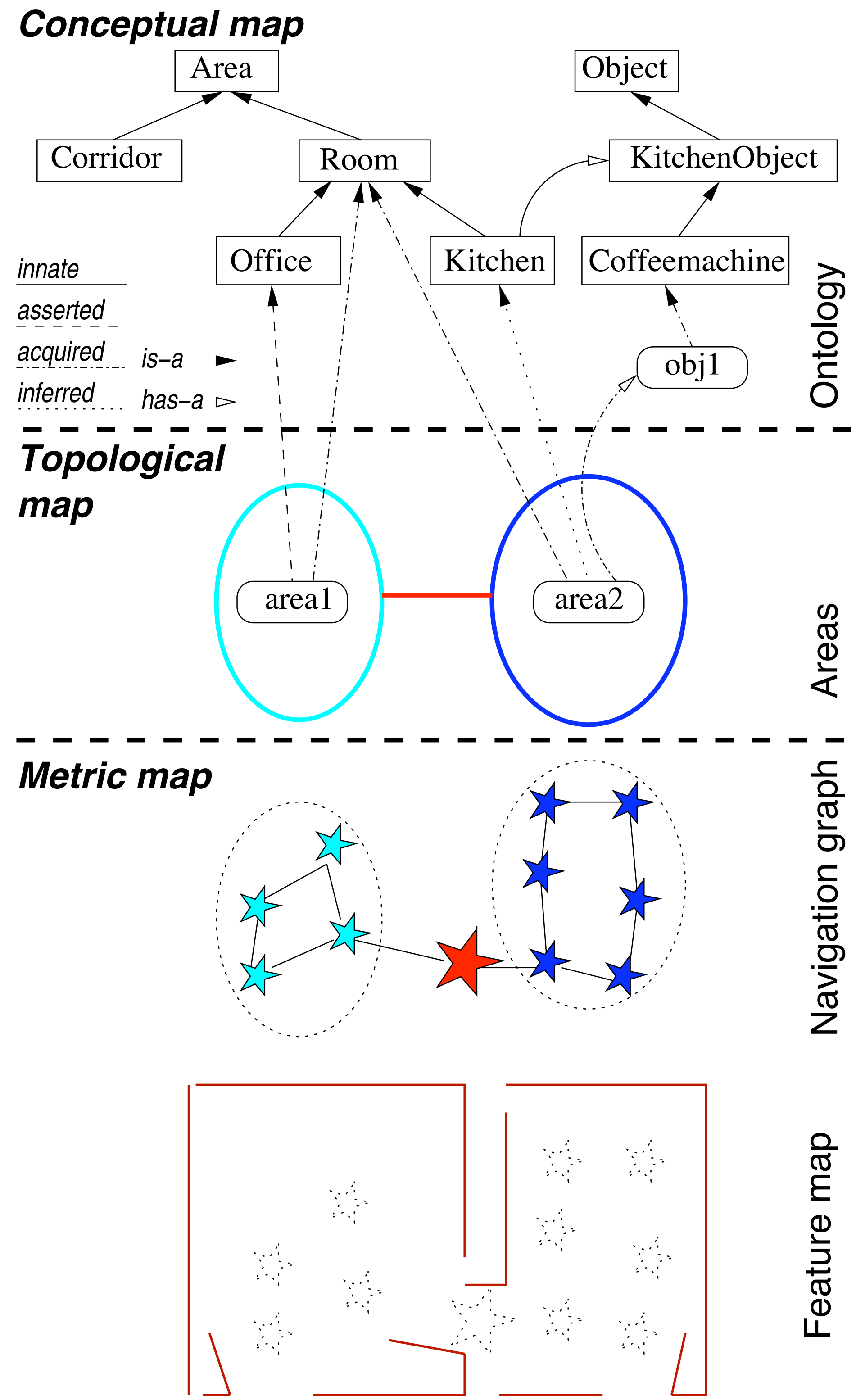


Example: Mapping



Example: Mapping

@{l1:area}(**living-room**)

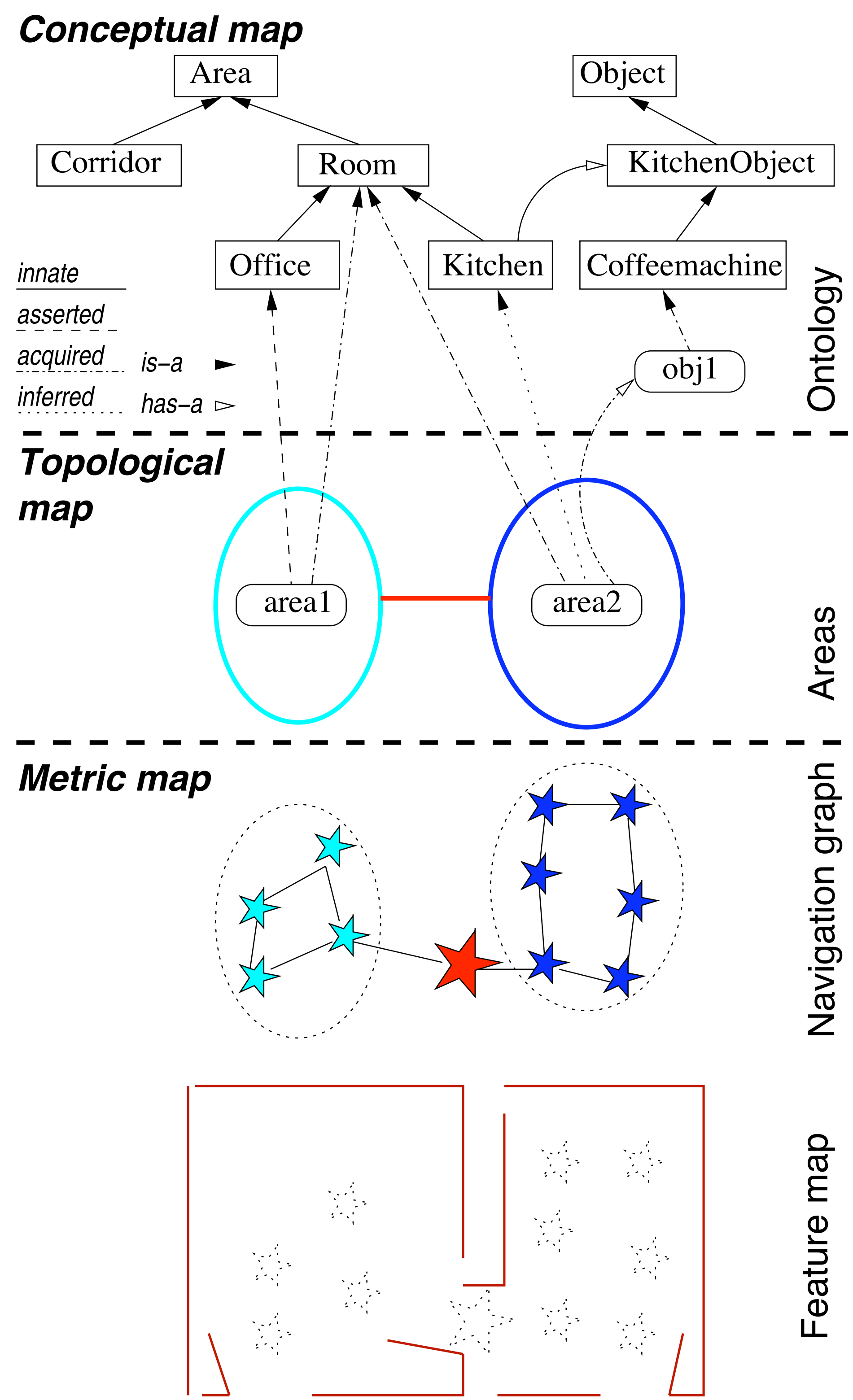


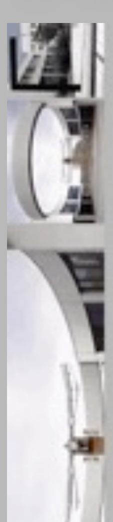
Example: Mapping

$$@\{l1{:}area\}(\textbf{living-room})$$

$$\updownarrow$$

$$\textit{concept}(\textbf{living-room})\ \&$$
$$\textit{instance}(L1,\textbf{living-room})\ \&$$
$$l1 \rightleftarrows L1$$





Example: Mapping

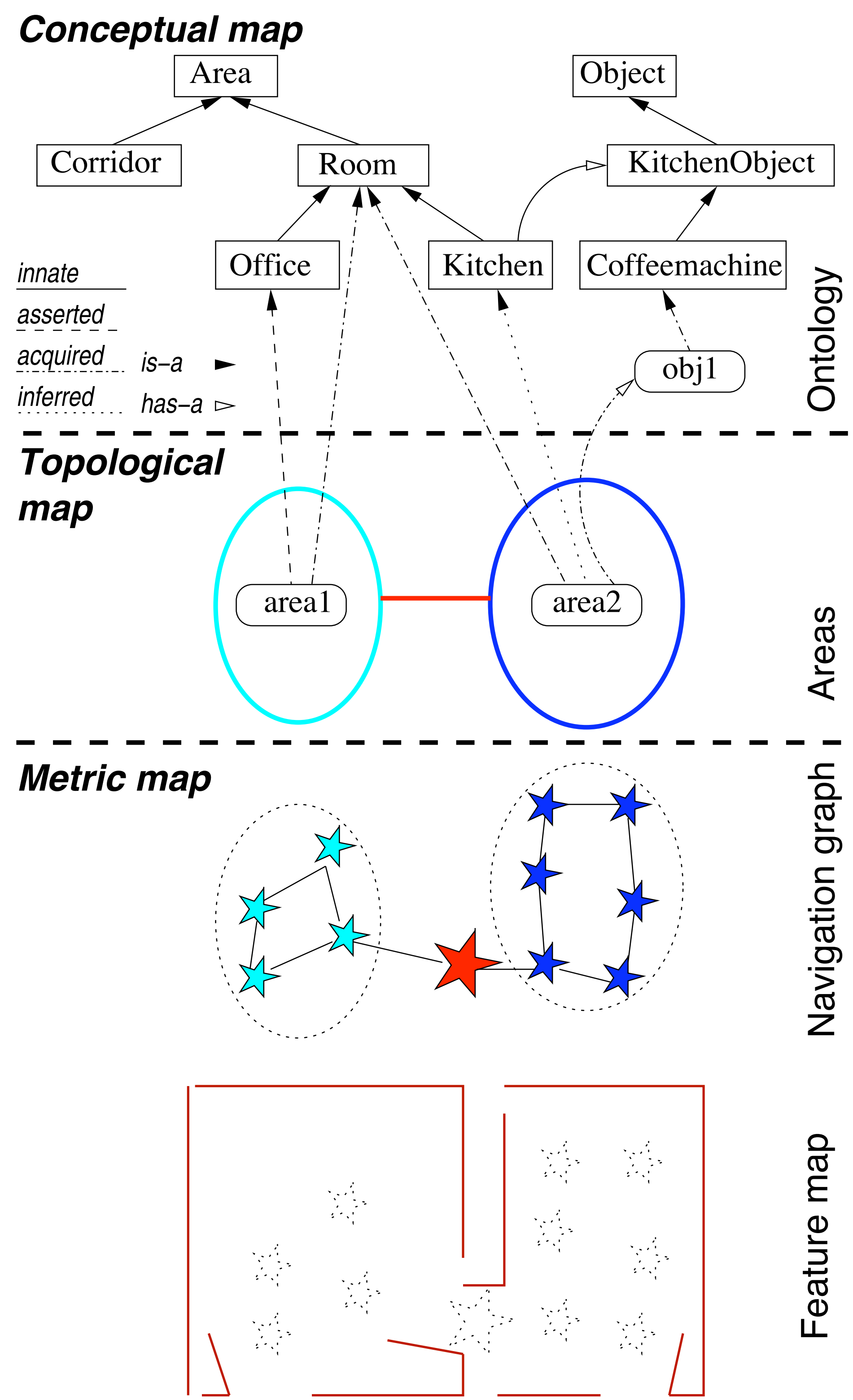
@{l1:area}(**living-room**)

↕

concept(**living-room**) & *instance*(L1,**living-room**) & l1 ⇄ L1

↕

area(area1) & area1 ⇄ L1



Example: Mapping (2005)

Example: Mapping (2005)

Example: Mapping (2005)

Example: Mapping (2005)

Example: Mapping (2005)

"Uncertainty"

Uncertainty in Mediation

- Ontology-based mediation
- Objects, features, uncertainty

Information Fusion For Visual Reference Resolution In Dynamic Situated Dialogue

Geert-Jan M. Kruijff[1], John D. Kelleher[2], and Nick Hawes[3]

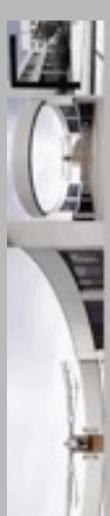
[1] Language Technology Lab, DFKI GmbH
gj@dfki.de,
WWW home page: http://www.dfki.de/~gj
[2] Dublin Institute of Technology
John.Kelleher@comp.dit.ie,
WWW home page: www.computing.dcu.ie/jkelleher/
[3] School of Computer Science, University of Birmingham
n.a.hawes@cs.bham.ac.uk,
WWW home page: http://www.cs.bham.ac.uk/~nah

Abstract. Human-Robot Interaction (HRI) invariably involves dialogue about objects in the environment in which the agents are situated. The paper focuses on the issue of resolving discourse references to such visual objects. The paper addresses the problem using strategies for *intra-modal fusion* (identifying that different occurrences concern the same object), and *inter-modal fusion*, (relating object references across different modalities). Core to these strategies are sensori-motoric coordination, and ontology-based mediation between content in different modalities. The approach has been fully implemented, and is illustrated with several working examples.

1 Introduction

The context of this work is the development of dialog systems for human-robot collaboration. The framework presented in this paper addresses a particular aspect of situated dialog, namely reference resolution. Reference resolution in situated dialog is a particular instance of the anchoring problem [Coradeschi and Saffiotti, 2003]: how can an artificial system create and maintain correspondences between the symbols and sensor data that refer to the same physical object?

In a dialog, human participants expect their partner to construct and maintain a model of the evolving linguistic context. Each referring expression used in the dialog introduces a representation into the semantics of its utterance. This representation must be bound to an element in the context model in order for the utterance's semantics to be fully resolved. Referring expressions that access a representation in the context are called *anaphoric*. In a *situated* dialog, human participants expect their partner to not only construct and maintain a model of the linguistic discourse, but also to have full perceptual knowledge of the environment. This introduces a form of reference, called *exophoric* reference. Exophoric references denote objects that have entered the dialog context through a non-linguistic modality (such as vision), but have not been previously evoked into the context. Consequently, for a robot to participate in a situated dialog,



Uncertainty in Mediation

- Ontology-based mediation
- Objects, features, uncertainty

Crossmodal Content Binding in Information-Processing Architectures*

Henrik Jacobsson
henrikj@dfki.de

Nick Hawes
n.a.hawes@cs.bham.ac.uk

Geert-Jan Kruijff
Language Technology Lab,
DFKI GmbH, Germany
gj@dfki.de

Jeremy Wyatt
School of Computer Science,
University of Birmingham, UK
j.l.wyatt@cs.bham.ac.uk

ABSTRACT

Operating in a physical context, an intelligent robot faces two fundamental problems. First, it needs to combine information from its different sensors to form a representation of the environment that is more complete than any of its sensors on its own could provide. Second, it needs to combine high-level representations (such as those for planning and dialogue) with its sensory information, to ensure that the interpretations of these symbolic representations are grounded in the situated context. Previous approaches to this problem have used techniques such as (low-level) information fusion, ontological reasoning, and (high-level) concept learning. This paper presents a framework in which these, and other approaches, can be combined to form a shared representation of the current state of the robot in relation to its environment and other agents. Preliminary results from an implemented system are presented to illustrate how the framework supports behaviours commonly required of an intelligent robot.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics

General Terms

Algorithms, Design

1. INTRODUCTION

An information-processing architecture for robotics is typically composed of a large number of cooperating subsystems, such as natural language analysis and production, computer vision, motoric skills, and various deliberative processes such as symbolic planners. The challenge addressed in this paper is the production and maintenance of a model of the world for a robot situated in "everyday" scenarios involving human interaction. This requires a method for *binding* representations across the subsystems. This world model should adequately reflect the aspects of the world that are stable in the medium term, whilst incorporating more dynamic aspects.

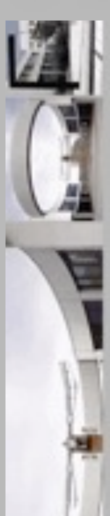
Throughout this paper we will primarily consider a robot that can interact with a human and a set of objects on a tabletop. For example, when faced with a scene containing a red mug, a blue cup and a blue bowl, the robot may be asked to "put the blue things to the left of the red thing". For a system to be able to perform such a task effectively, it must be able to build a representation that connects the (low-level and modality specific) information about the world and the (high-level and amodal) representations that can be used to interpret the utterance, determine the desired world state, and plan behaviour. As resulting actions must be executed in the world, the representation must allow the robot to ultimately access the low-level (i.e. metric) information from which its higher-level representations are derived.

Any design for a system to tackle the above task must focus on creating such a representation, and grounding it in the environment of the robot. In addition to this, the engineering effort of integrating the various information-processing subsystems with the representation must be considered. After all, since the robot is an engineered system, every component must be put there by means of human effort.

The grounding problem is entangled with the engineering problem of subsystem integration and cannot be considered in isolation. Grounding can generally be seen as the process of establishing the relation between a representation in one domain with that of another. One special case is is when one of the domains is the external world, i.e. "reality":

> The term grounding [denotes] the processes by which an agent relates beliefs to external physical objects. Agents use grounding processes to construct models of, predict, and react to, their external environment. Language grounding refers to processes specialised for relating words and speech acts to a language user's environment via grounded beliefs. [11] p. 8

*This work was supported by the EU FP6 IST Cognitive Systems Integrated Project "CoSy" FP6-004250-IP.



Uncertainty in Mediation

- Ontology-based mediation
- Objects, features, uncertainty

Conceptual meaning:
concept(**box**) & *instance*(I1,**box**) & i1 ⇄ I1

↕

Mapping

↕

Linguistic meaning:
@{i1:object}(**box**)

Crossmodal Content Binding in Information-Processing Architectures*

Henrik Jacobsson
henrikj@dfki.de

Nick Hawes
n.a.hawes@cs.bham.ac.uk

Geert-Jan Kruijff
Language Technology Lab,
DFKI GmbH, Germany
gj@dfki.de

Jeremy Wyatt
School of Computer Science,
University of Birmingham, UK
j.l.wyatt@cs.bham.ac.uk

ABSTRACT

Operating in a physical context, an intelligent robot faces two fundamental problems. First, it needs to combine information from its different sensors to form a representation of the environment that is more complete than any of its sensors on its own could provide. Second, it needs to combine high-level representations (such as those for planning and dialogue) with its sensory information, to ensure that the interpretations of these symbolic representations are grounded in the situated context. Previous approaches to this problem have used techniques such as (low-level) information fusion, ontological reasoning, and (high-level) concept learning. This paper presents a framework in which these, and other approaches, can be combined to form a shared representation of the current state of the robot in relation to its environment and other agents. Preliminary results from an implemented system are presented to illustrate how the framework supports behaviours commonly required of an intelligent robot.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics

General Terms

Algorithms, Design

1. INTRODUCTION

An information-processing architecture for robotics is typically composed of a large number of cooperating subsystems, such as natural language analysis and production, computer vision, motoric skills, and various deliberative processes such as symbolic planners. The challenge addressed in this paper is the production and maintenance of a model of the world for a robot situated in "everyday" scenarios involving human interaction. This requires a method for *binding* representations across the subsystems. This world model should adequately reflect the aspects of the world that are stable in the medium term, whilst incorporating more dynamic aspects.

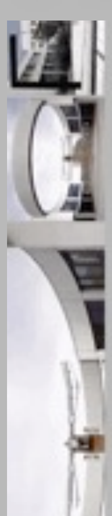
Throughout this paper we will primarily consider a robot that can interact with a human and a set of objects on a tabletop. For example, when faced with a scene containing a red mug, a blue cup and a blue bowl, the robot may be asked to "put the blue things to the left of the red thing". For a system to be able to perform such a task effectively, it must be able to build a representation that connects the (low-level and modality specific) information about the world and the (high-level and amodal) representations that can be used to interpret the utterance, determine the desired world state, and plan behaviour. As resulting actions must be executed in the world, the representation must allow the robot to ultimately access the low-level (i.e. metric) information from which its higher-level representations are derived.

Any design for a system to tackle the above task must focus on creating such a representation, and grounding it in the environment of the robot. In addition to this, the engineering effort of integrating the various information-processing subsystems with the representation must be considered. After all, since the robot is an engineered system, every component must be put there by means of human effort.

The grounding problem is entangled with the engineering problem of subsystem integration and cannot be considered in isolation. Grounding can generally be seen as the process of establishing the relation between a representation in one domain with that of another. One special case is is when one of the domains is the external world, i.e. "reality":

> The term grounding [denotes] the processes by which an agent relates beliefs to external physical objects. Agents use grounding processes to construct models of, predict, and react to, their external environment. Language grounding refers to processes specialised for relating words and speech acts to a language user's environment via grounded beliefs. [11] p. 8

*This work was supported by the EU FP6 IST Cognitive Systems Integrated Project "CoSy" FP6-004250-IP.

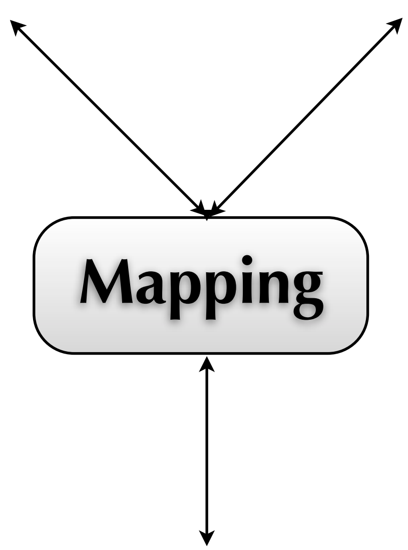


Uncertainty in Mediation

- Ontology-based mediation
- Objects, features, uncertainty

Conceptual meaning: *concept*(**box**) & ...

Modal meaning: *c1*(**box**) & f1(...) ..



Linguistic meaning: @{i1:object}(**box**)

Crossmodal Content Binding in Information-Processing Architectures*

Henrik Jacobsson
henrikj@dfki.de

Nick Hawes
n.a.hawes@cs.bham.ac.uk

Geert-Jan Kruijff
Language Technology Lab,
DFKI GmbH, Germany
gj@dfki.de

Jeremy Wyatt
School of Computer Science,
University of Birmingham, UK
j.l.wyatt@cs.bham.ac.uk

ABSTRACT

Operating in a physical context, an intelligent robot faces two fundamental problems. First, it needs to combine information from its different sensors to form a representation of the environment that is more complete than any of its sensors on its own could provide. Second, it needs to combine high-level representations (such as those for planning and dialogue) with its sensory information, to ensure that the interpretations of these symbolic representations are grounded in the situated context. Previous approaches to this problem have used techniques such as (low-level) information fusion, ontological reasoning, and (high-level) concept learning. This paper presents a framework in which these, and other approaches, can be combined to form a shared representation of the current state of the robot in relation to its environment and other agents. Preliminary results from an implemented system are presented to illustrate how the framework supports behaviours commonly required of an intelligent robot.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics

General Terms

Algorithms, Design

1. INTRODUCTION

An information-processing architecture for robotics is typically composed of a large number of cooperating subsystems, such as natural language analysis and production, computer vision, motoric skills, and various deliberative processes such as symbolic planners. The challenge addressed in this paper is the production and maintenance of a model of the world for a robot situated in "everyday" scenarios involving human interaction. This requires a method for *binding* representations across the subsystems. This world model should adequately reflect the aspects of the world that are stable in the medium term, whilst incorporating more dynamic aspects.

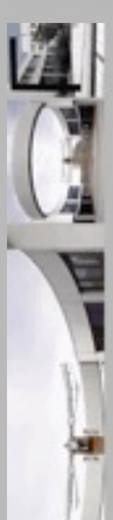
Throughout this paper we will primarily consider a robot that can interact with a human and a set of objects on a tabletop. For example, when faced with a scene containing a red mug, a blue cup and a blue bowl, the robot may be asked to "put the blue things to the left of the red thing". For a system to be able to perform such a task effectively, it must be able to build a representation that connects the (low-level and modality specific) information about the world and the (high-level and amodal) representations that can be used to interpret the utterance, determine the desired world state, and plan behaviour. As resulting actions must be executed in the world, the representation must allow the robot to ultimately access the low-level (i.e. metric) information from which its higher-level representations are derived.

Any design for a system to tackle the above task must focus on creating such a representation, and grounding it in the environment of the robot. In addition to this, the engineering effort of integrating the various information-processing subsystems with the representation must be considered. After all, since the robot is an engineered system, every component must be put there by means of human effort.

The grounding problem is entangled with the engineering problem of subsystem integration and cannot be considered in isolation. Grounding can generally be seen as the process of establishing the relation between a representation in one domain with that of another. One special case is is when one of the domains is the external world, i.e. "reality":

> The term grounding [denotes] the processes by which an agent relates beliefs to external physical objects. Agents use grounding processes to construct models of, predict, and react to, their external environment. Language grounding refers to processes specialised for relating words and speech acts to a language user's environment via grounded beliefs. [11] p. 8

*This work was supported by the EU FP6 IST Cognitive Systems Integrated Project "CoSy" FP6-004250-IP.

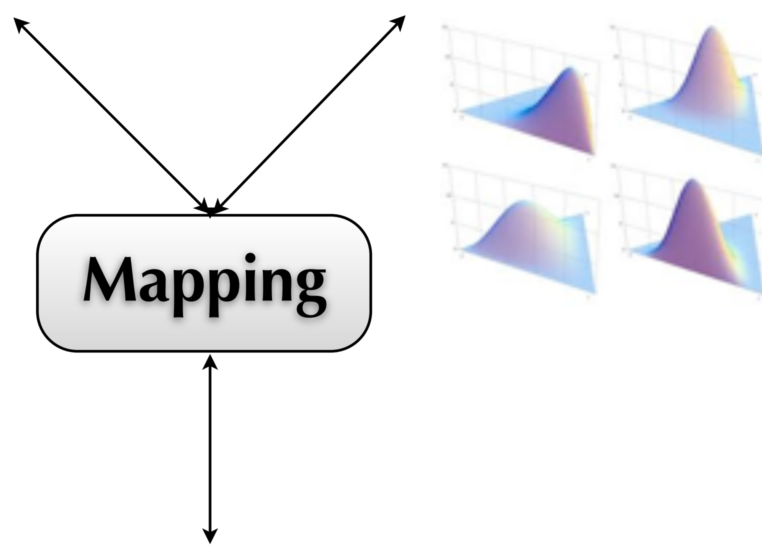


Uncertainty in Mediation

- Ontology-based mediation
- Objects, features, uncertainty

Conceptual meaning: *concept*(**box**) & ...

Modal meaning: *c1*(**box**) & f1(...) ..



Linguistic meaning: @{i1:object}(**box**)

Crossmodal Content Binding in Information-Processing Architectures*

Henrik Jacobsson
henrikj@dfki.de

Nick Hawes
n.a.hawes@cs.bham.ac.uk

Geert-Jan Kruijff
Language Technology Lab,
DFKI GmbH, Germany
gj@dfki.de

Jeremy Wyatt
School of Computer Science,
University of Birmingham, UK
j.l.wyatt@cs.bham.ac.uk

ABSTRACT

Operating in a physical context, an intelligent robot faces two fundamental problems. First, it needs to combine information from its different sensors to form a representation of the environment that is more complete than any of its sensors on its own could provide. Second, it needs to combine high-level representations (such as those for planning and dialogue) with its sensory information, to ensure that the interpretations of these symbolic representations are grounded in the situated context. Previous approaches to this problem have used techniques such as (low-level) information fusion, ontological reasoning, and (high-level) concept learning. This paper presents a framework in which these, and other approaches, can be combined to form a shared representation of the current state of the robot in relation to its environment and other agents. Preliminary results from an implemented system are presented to illustrate how the framework supports behaviours commonly required of an intelligent robot.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics

General Terms

Algorithms, Design

1. INTRODUCTION

An information-processing architecture for robotics is typically composed of a large number of cooperating subsystems, such as natural language analysis and production, computer vision, motoric skills, and various deliberative processes such as symbolic planners. The challenge addressed in this paper is the production and maintenance of a model of the world for a robot situated in "everyday" scenarios involving human interaction. This requires a method for *binding* representations across the subsystems. This world model should adequately reflect the aspects of the world that are stable in the medium term, whilst incorporating more dynamic aspects.

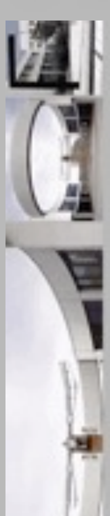
Throughout this paper we will primarily consider a robot that can interact with a human and a set of objects on a tabletop. For example, when faced with a scene containing a red mug, a blue cup and a blue bowl, the robot may be asked to "put the blue things to the left of the red thing". For a system to be able to perform such a task effectively, it must be able to build a representation that connects the (low-level and modality specific) information about the world and the (high-level and amodal) representations that can be used to interpret the utterance, determine the desired world state, and plan behaviour. As resulting actions must be executed in the world, the representation must allow the robot to ultimately access the low-level (i.e. metric) information from which its higher-level representations are derived.

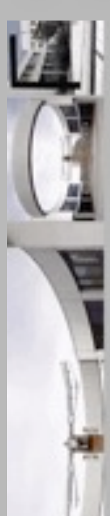
Any design for a system to tackle the above task must focus on creating such a representation, and grounding it in the environment of the robot. In addition to this, the engineering effort of integrating the various information-processing subsystems with the representation must be considered. After all, since the robot is an engineered system, every component must be put there by means of human effort.

The grounding problem is entangled with the engineering problem of subsystem integration and cannot be considered in isolation. Grounding can generally be seen as the process of establishing the relation between a representation in one domain with that of another. One special case is is when one of the domains is the external world, i.e. "reality":

> The term grounding [denotes] the processes by which an agent relates beliefs to external physical objects. Agents use grounding processes to construct models of, predict, and react to, their external environment. Language grounding refers to processes specialised for relating words and speech acts to a language user's environment via grounded beliefs. [11] p. 8

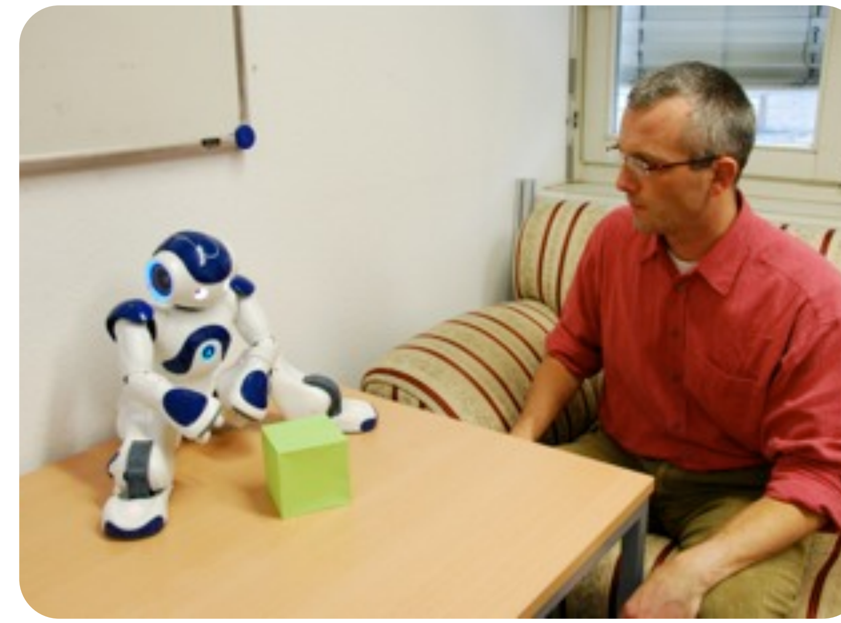
*This work was supported by the EU FP6 IST Cognitive Systems Integrated Project "CoSy" FP6-004250-IP.

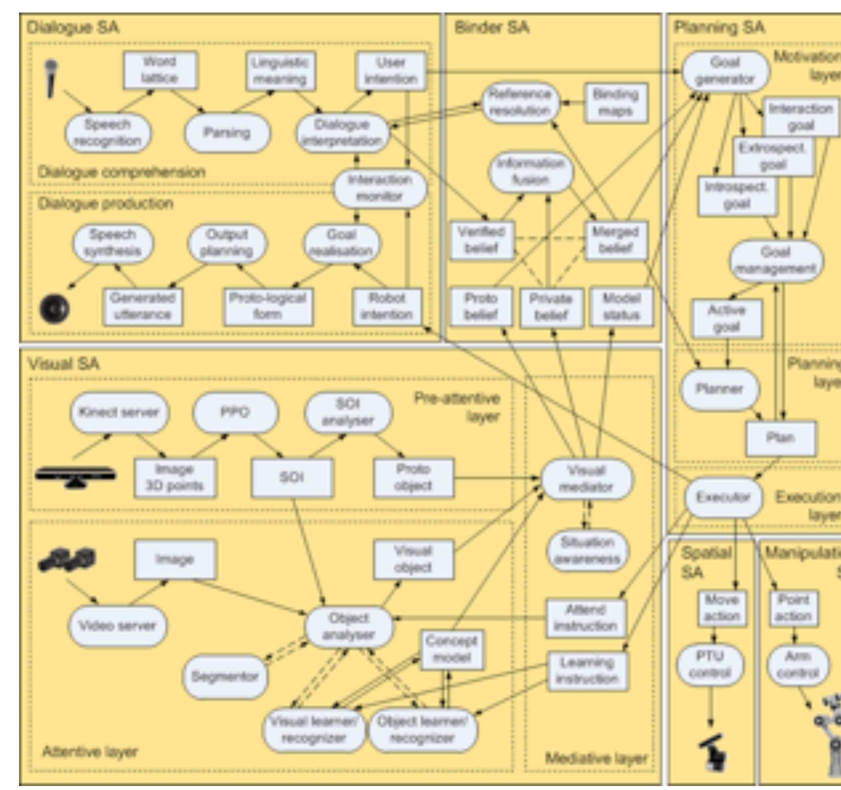




Example: Interactive learning of visual objects

- Uncertainty in visual classification
- Uncertainty in mapping visual categories and linguistic meaning
- Incremental, “life-long learning” process to reduce uncertainty
- Key: clarification,confirmation





EXAMPLE: INTERACTIVE LEARNING OF VISUAL OBJECTS

- Uncertainty in visual classification
- Uncertainty in mapping visual categories and linguistic meaning
- Incremental, "life-long learning" process to reduce uncertainty
- Key: clarification,confirmation

A system for interactive learning in dialogue with a tutor

Danijel Skočaj, Matej Kristan, Alen Vrečko, Marko Mahnič, Miroslav Janíček, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich and Kai Zhou

Abstract— **In this paper we present representations and mechanisms that facilitate continuous learning of visual concepts in dialogue with a tutor and show the implemented robot system. We present how beliefs about the world are created by processing visual and linguistic information and show how they are used for planning system behaviour with the aim at satisfying its internal drive – to extend its knowledge. The system facilitates different kinds of learning initiated by the human tutor or by the system itself. We demonstrate these principles in the case of learning about object colours and basic shapes.**

I. INTRODUCTION

Cognitive systems are often characterised by their ability to learn, communicate and act autonomously. By combining these competencies, the system can incrementally learn by engaging in mixed initiative dialogues with a human tutor. In this paper we focus on representations and mechanisms that enable such interactive learning and present a system designed to acquire visual concepts through interaction with a human.

Such continuous and interactive learning is important from several perspectives. A system operating in a real life environment is continuously exposed to new observations (scenes, objects, actions etc.) that cannot be envisioned in advance. Therefore, it has to be able to update its knowledge continuously based on the newly obtained visual information and information provided by a human teacher. Assuming that the information provided by the human is correct, such interactive learning can significantly facilitate, and increase the robustness of, the learning process, which is prone to errors due to unreliable robot perception capabilities. By assessing the system's knowledge, the human can adapt their way of teaching and drive the learning process more efficiently. Similarly, the robot can take the initiative, and ask the human for the information that would increase its knowledge most, which should in turn lead to more efficient learning.

In this paper we describe how our robot *George*, depicted in Fig. 1, learns and refines visual conceptual models of colours and two basic shapes, either by attending to information deliberately provided by a human tutor (*tutor-driven learning*: e.g., H: 'This is a red box.') or by taking initiative itself, asking the tutor for specific information about an object in the scene (*situated tutor-assisted learning*: e.g., G: 'Is the elongated object yellow?'), or even asking questions that are not related to the current scene (*non-situated tutor-assisted learning*: e.g., G: 'Can you show me something red?')[1]. Our approach unifies these cases into an integrated approach including incremental visual learning, selection of learning goals, continual planning to select actions for optimal learning behaviour, and a dialogue subsystem. George is one system in a family of integrated systems that aim to understand where their own knowledge is incomplete and that take actions to extend their knowledge subsequently. Our objective is to demonstrate that a cognitive system can efficiently acquire conceptual models in an interactive learning process that is not overly taxing with respect to tutor supervision and is performed in an intuitive, user-friendly way.

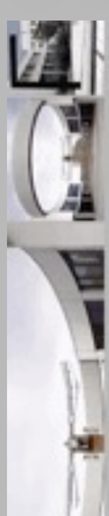
Fig. 1. Scenario setup.

Interactive continuous learning using information obtained from vision and language is a desirable property of any cognitive system, therefore several systems have been developed that address this issue (e.g., [1], [2], [3], [4], [5], [6], [7]). Different systems focus on different aspects of this problem, such as the system architecture and integration [3], [4], [6], learning [1], [2], [6], [7], or social interaction [5]. Our work focuses on the integration of visual perception and processing of linguistic information by forming beliefs about the state of the world; these beliefs are then used in the learning process for updating the current representations. The system behaviour is driven by a motivation framework which facilitates different kinds of learning in a dialogue with a human teacher, including self-motivated learning, triggered by autonomous knowledge gap detection. Also,

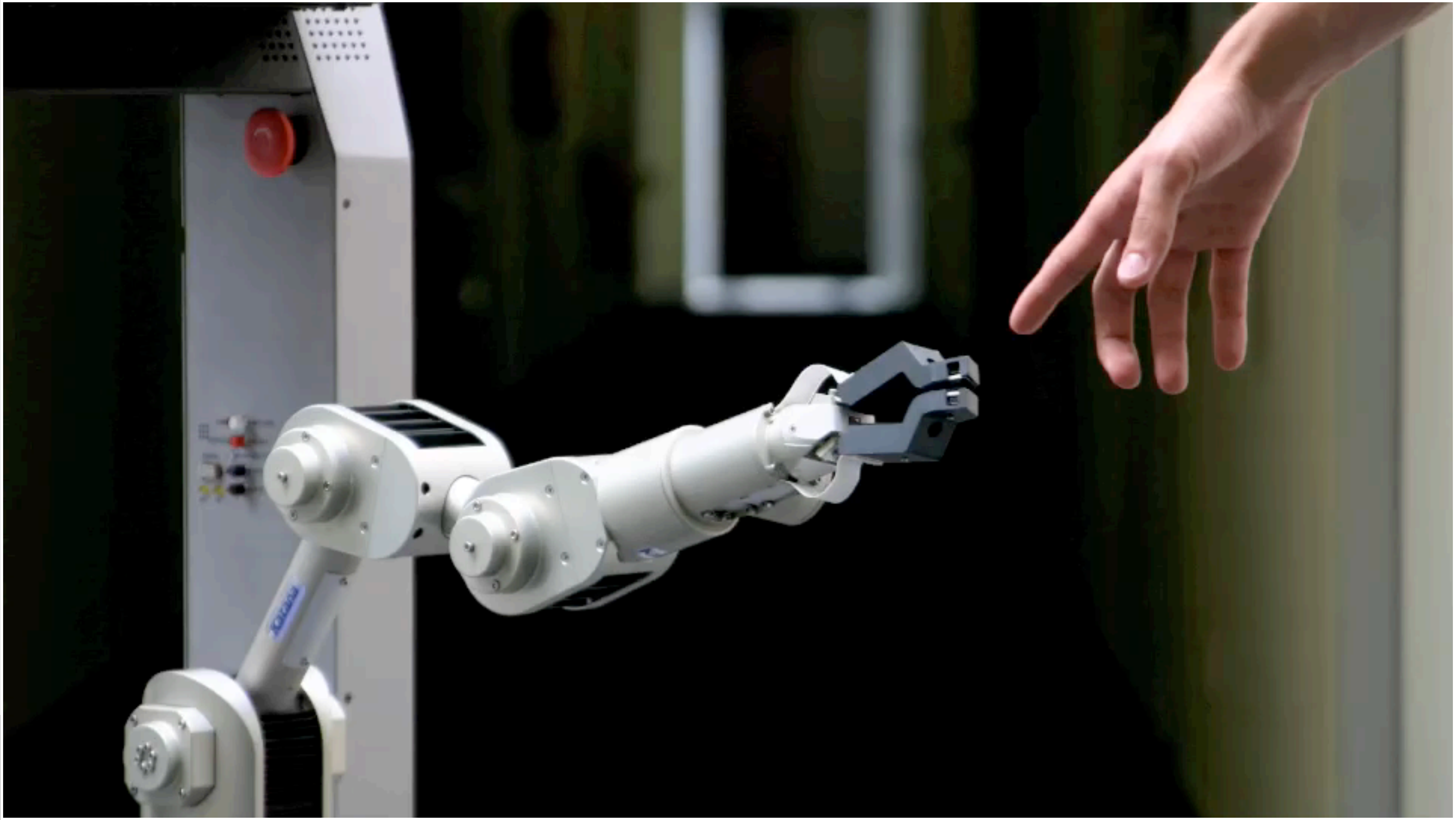
The work was supported by the EC FP7 IST project CogX-215181.
D. Skočaj, M. Kristan, A. Vrečko, and M. Mahnič are with University of Ljubljana, Slovenia
M. Janíček and G.J. M. Kruijff are with DFKI, Saarbrücken, Germany
M. Hanheide and N. Hawes are with University of Birmingham, UK
T. Keller is with Albert-Ludwigs-Universität Freiburg, Germany
M. Zillich and K. Zhou are with Vienna University of Technology, Austria

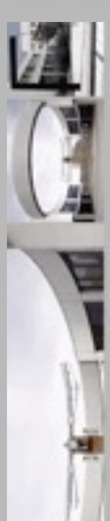
[1]The robot can be seen in action in the video accessible at http://cogx.eu/results/george.

Interactive learning (2009)



Interactive learning (2009)





VISUAL LEARNING

vision | binding | motivation | planning | dialogue

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify

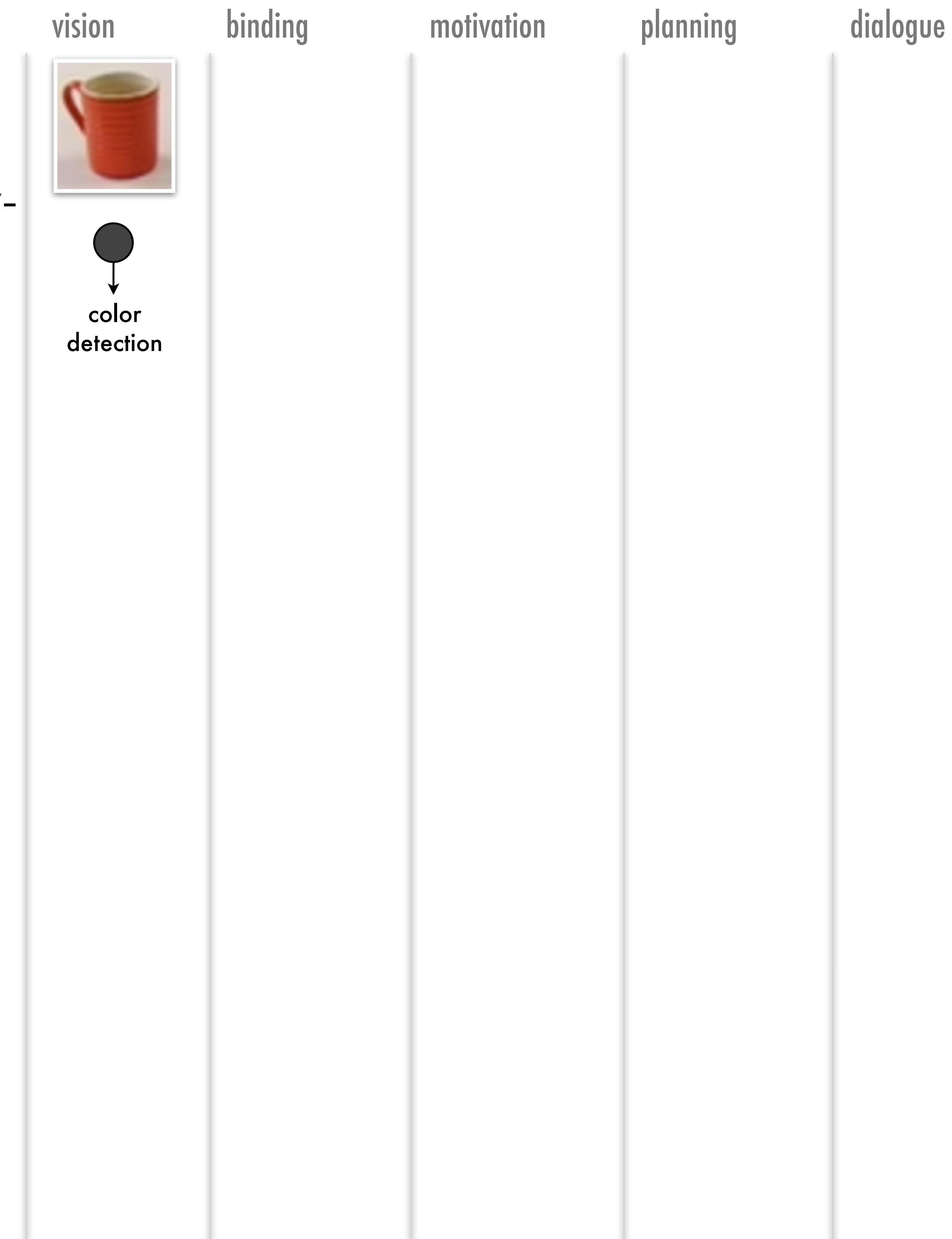
VISUAL LEARNING

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify

vision	binding	motivation	planning	dialogue
				

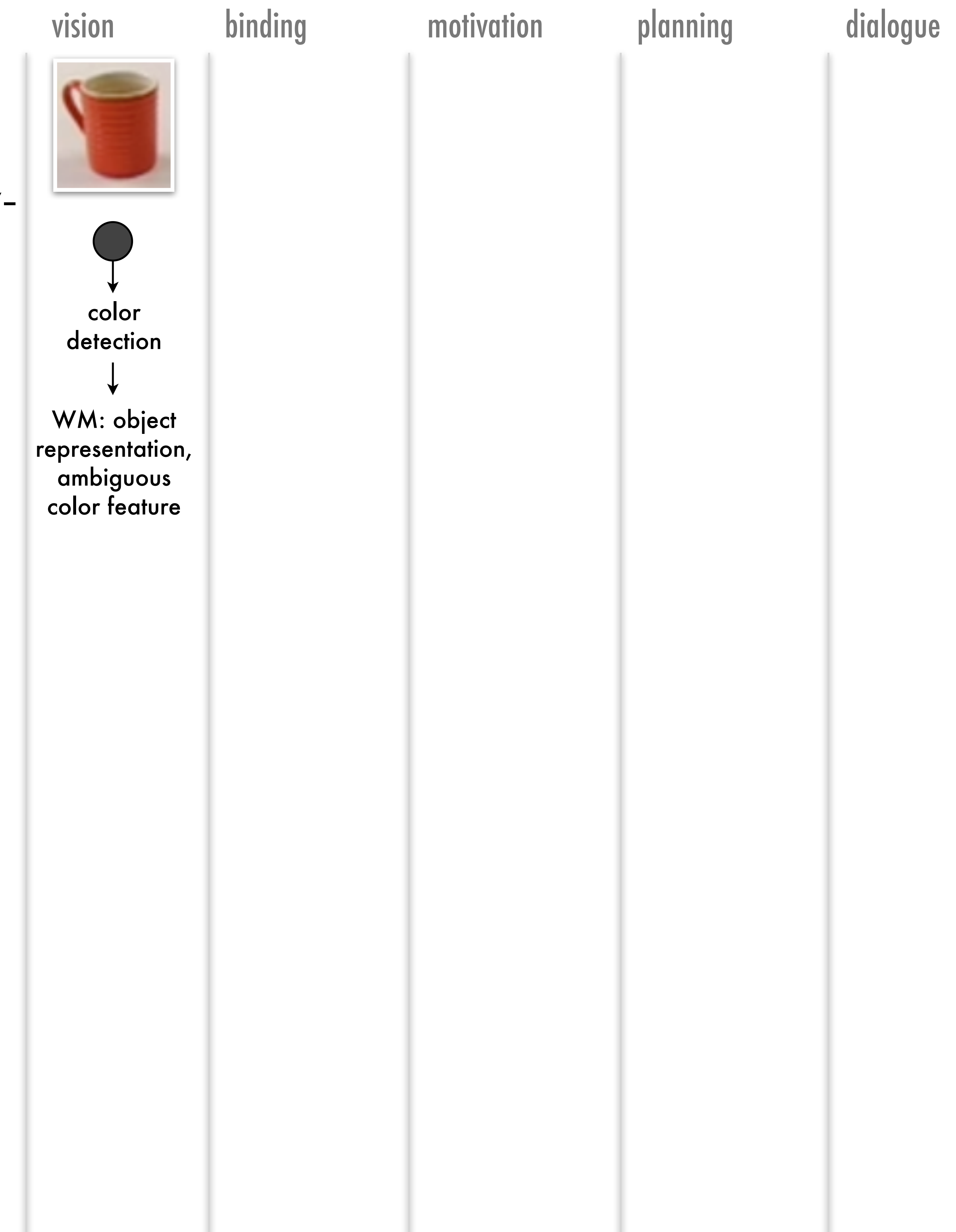
VISUAL LEARNING

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify



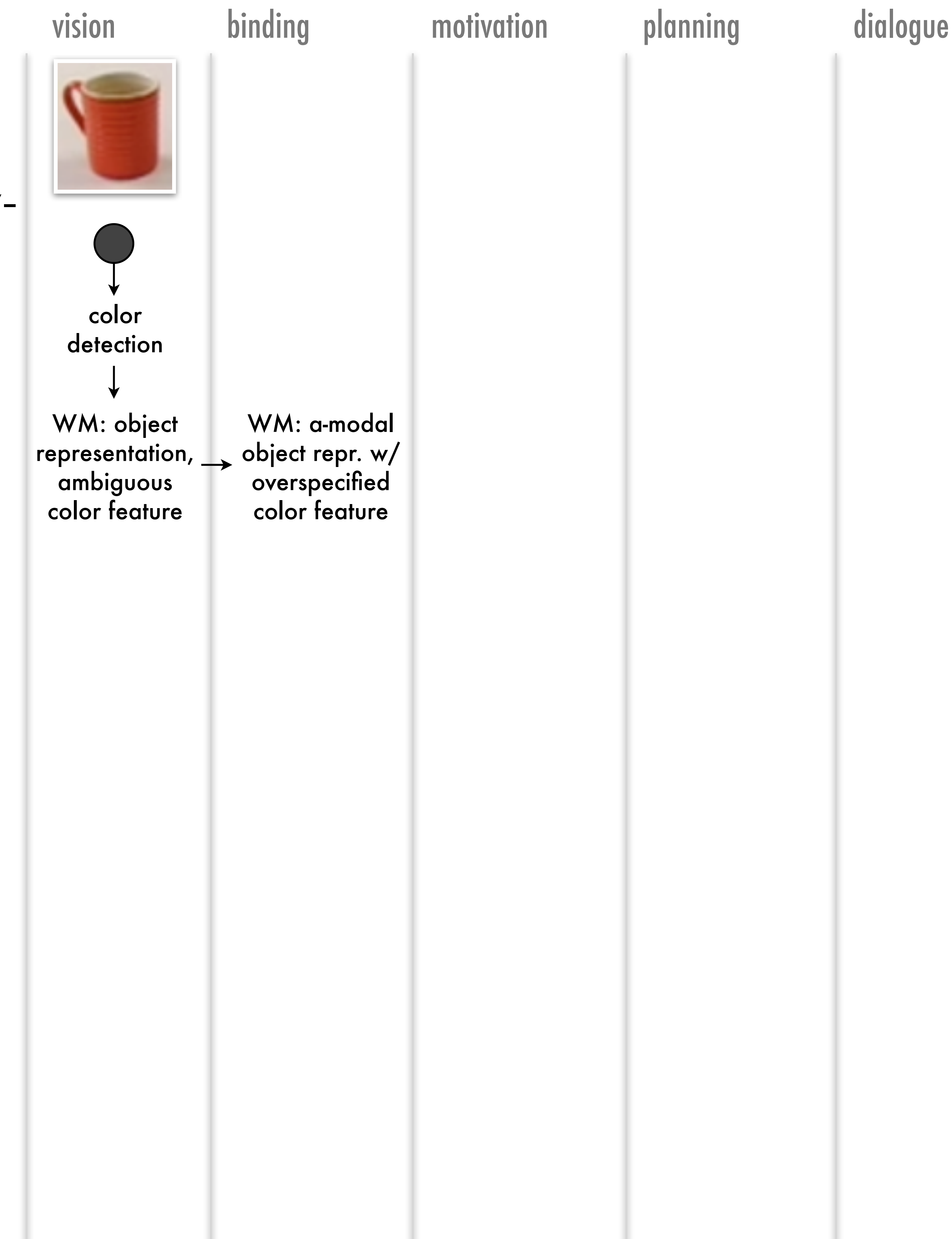
VISUAL LEARNING

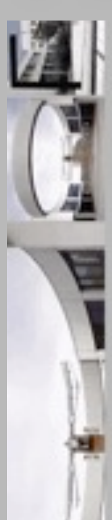
- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify



VISUAL LEARNING

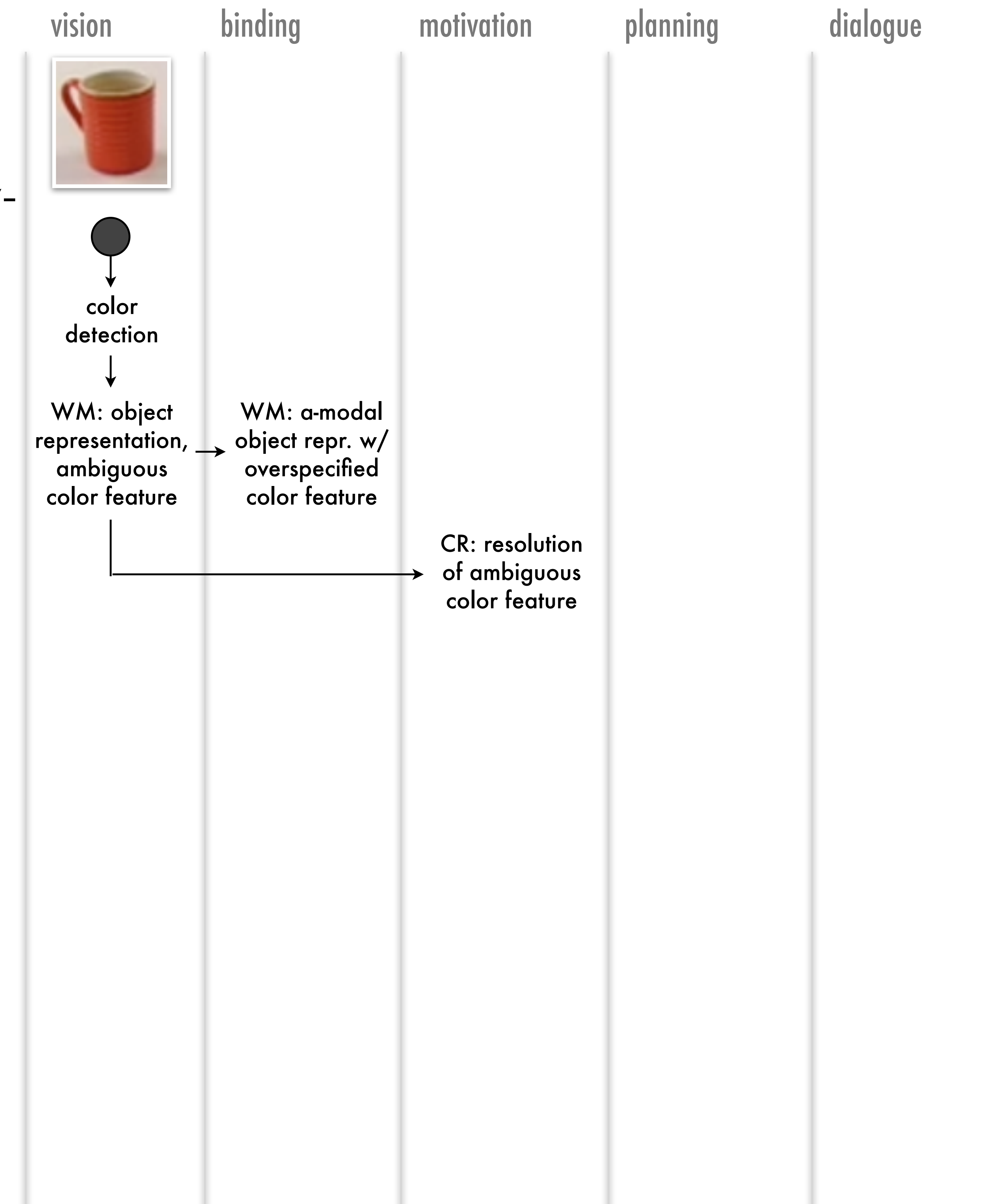
- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify





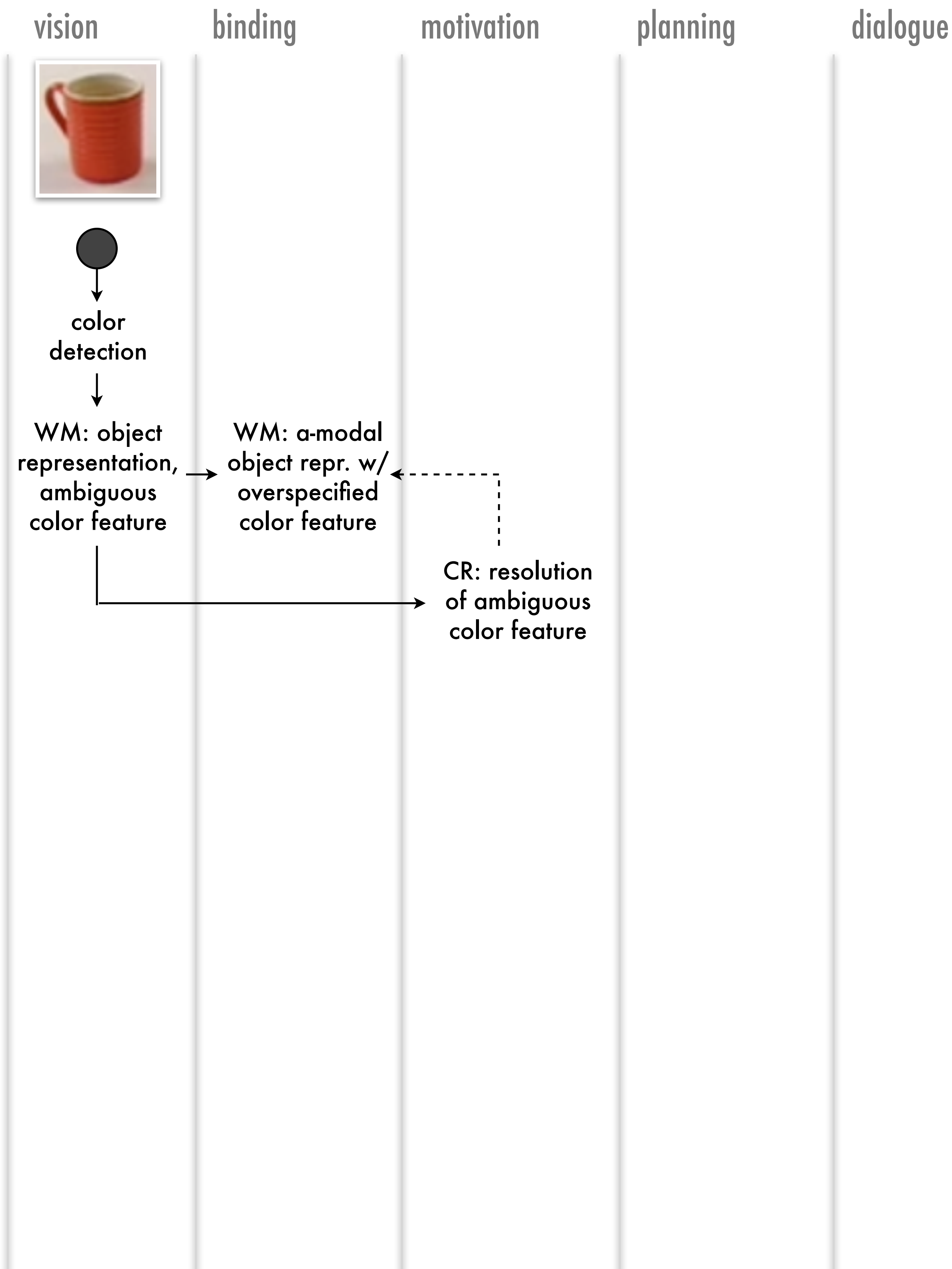
VISUAL LEARNING

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify



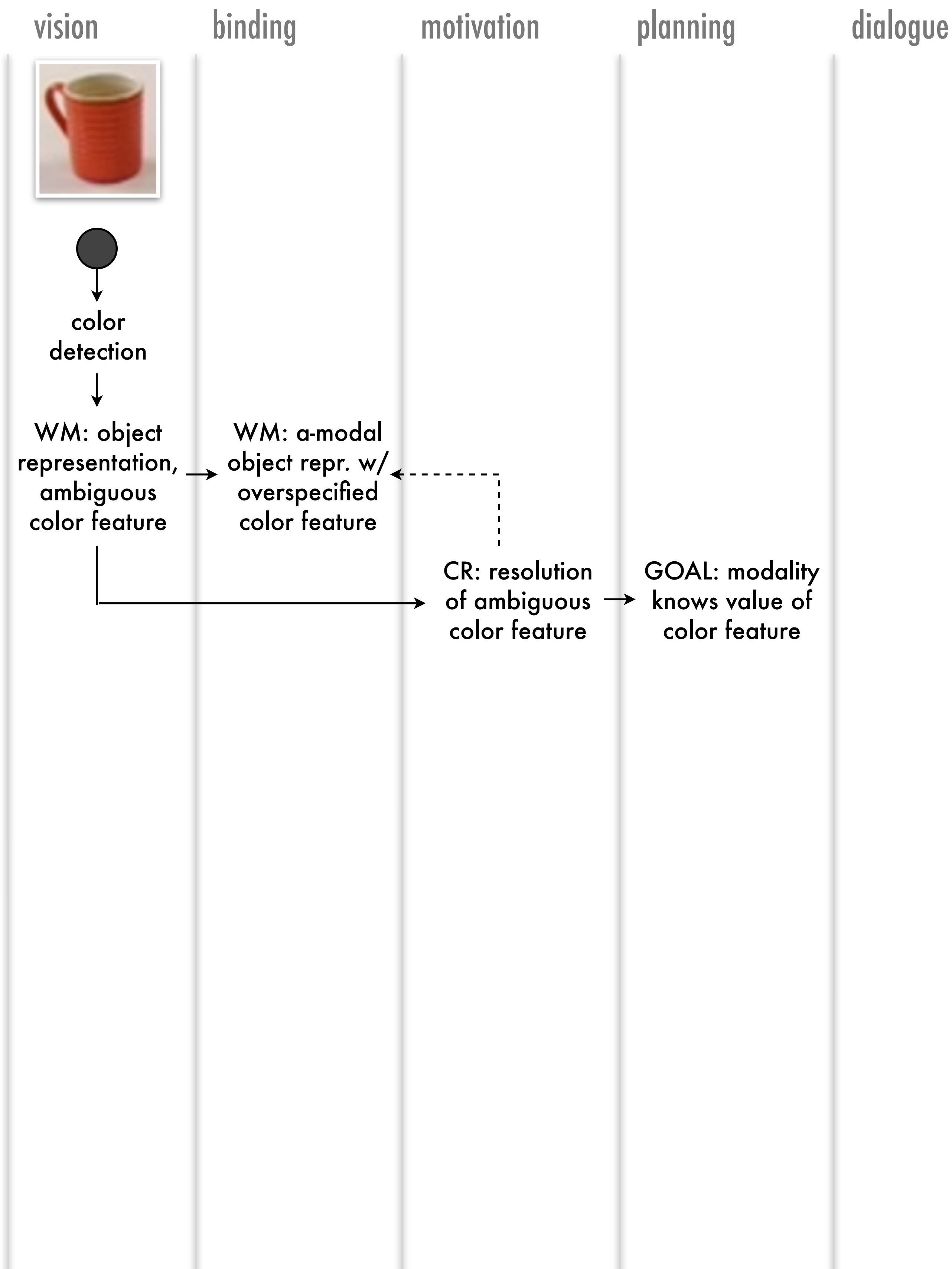
VISUAL LEARNING

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify



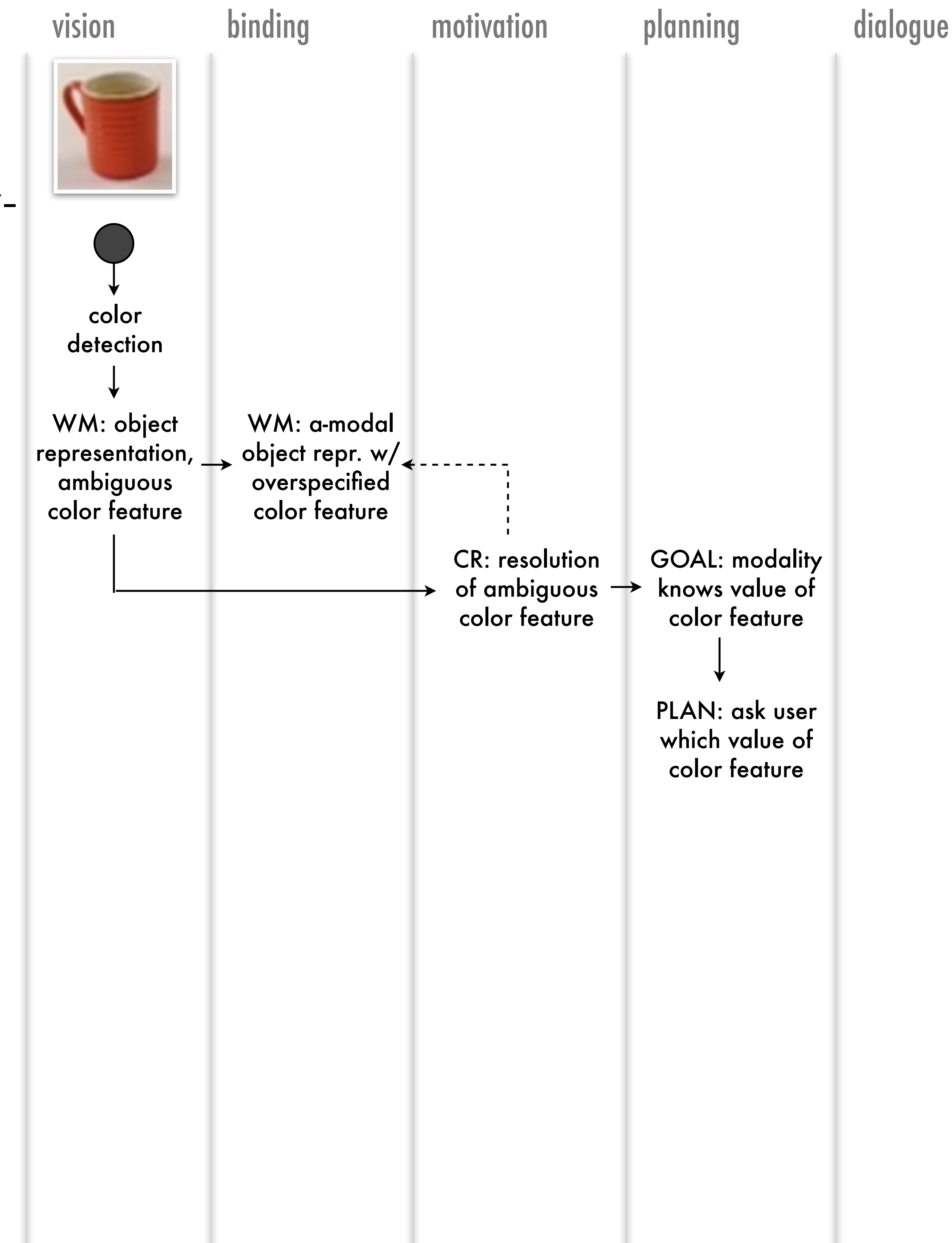
VISUAL LEARNING

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify



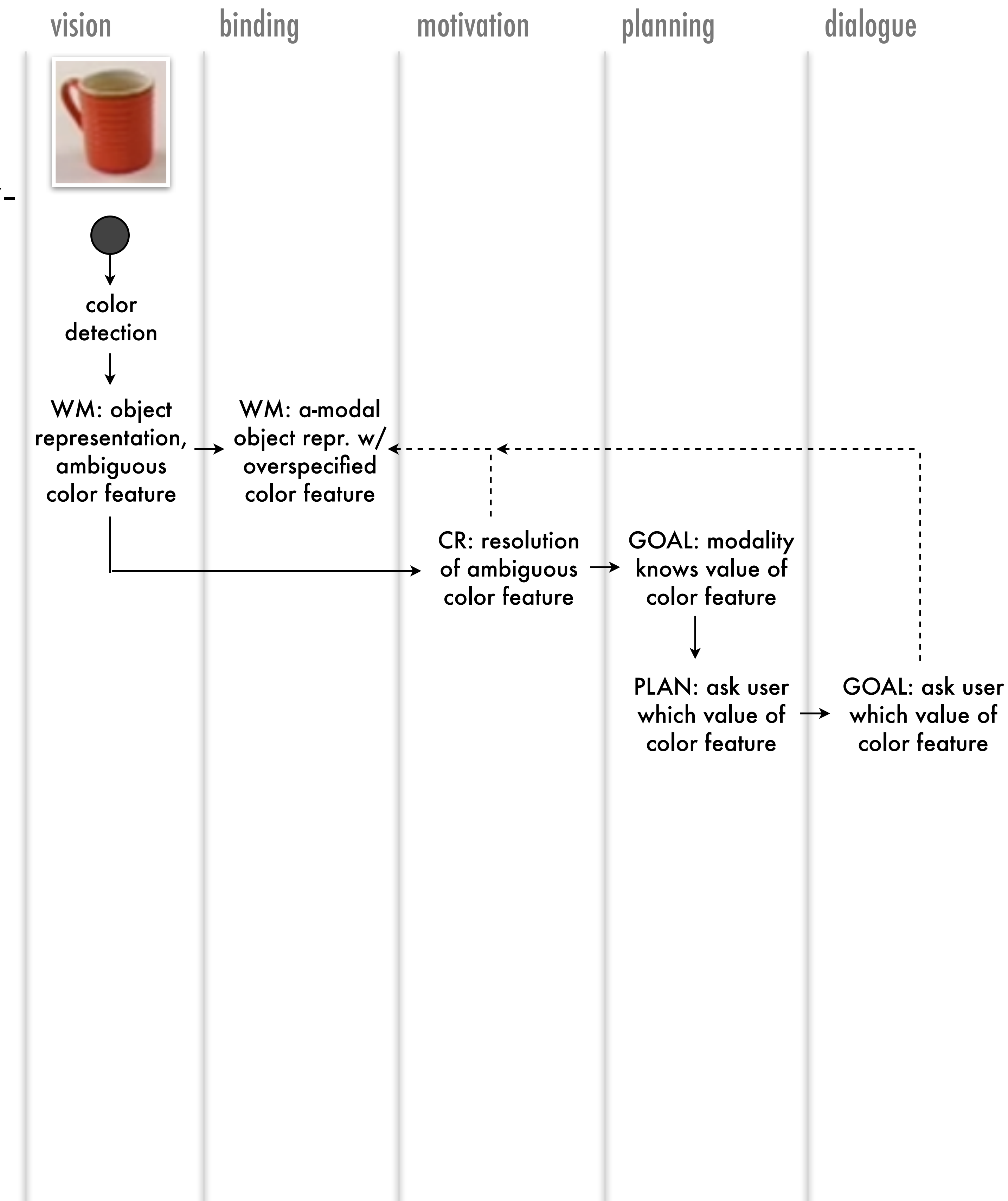
VISUAL LEARNING

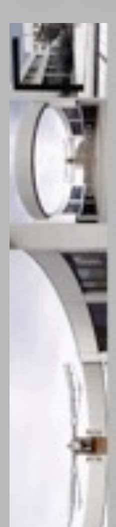
- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify



VISUAL LEARNING

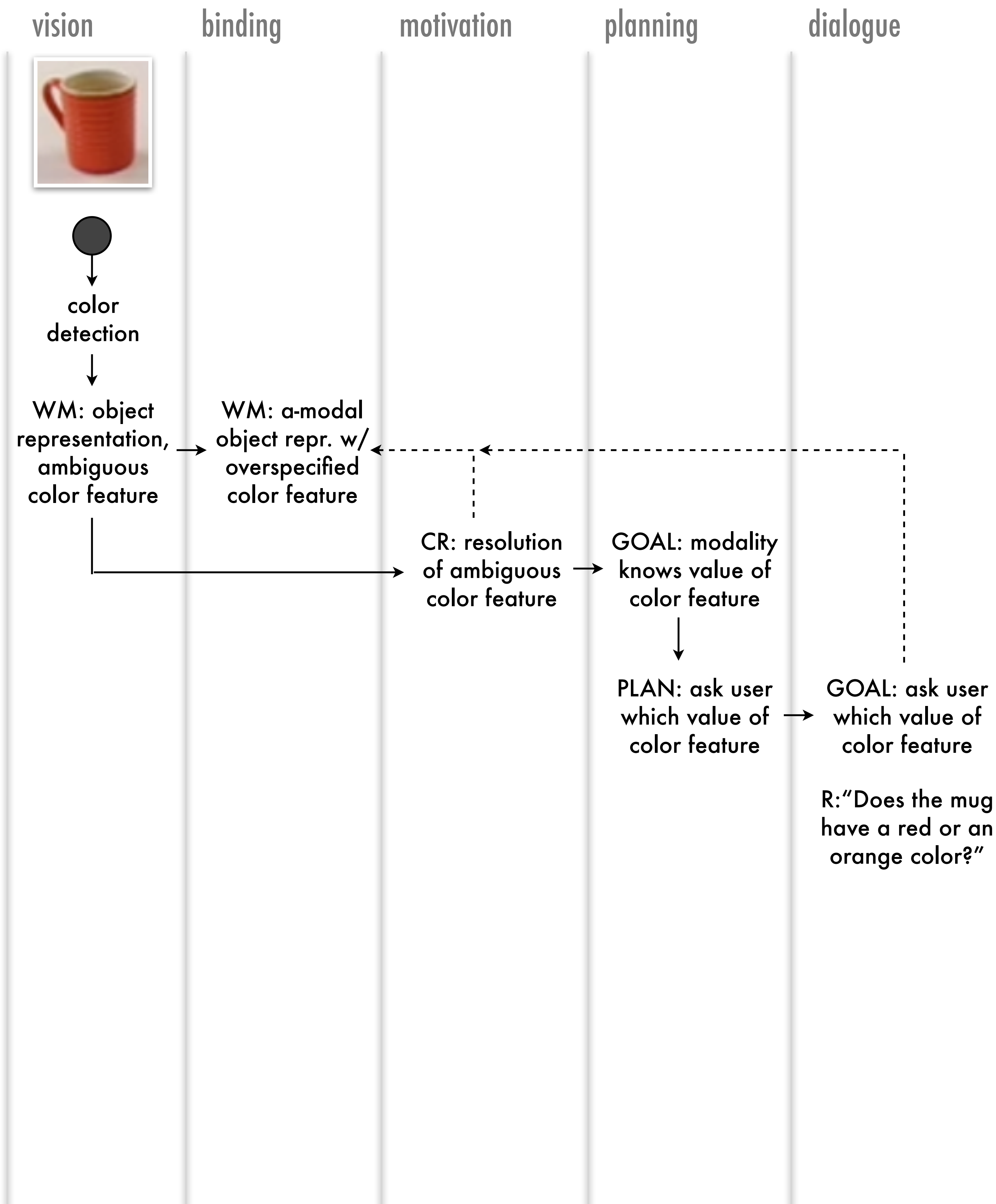
- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify





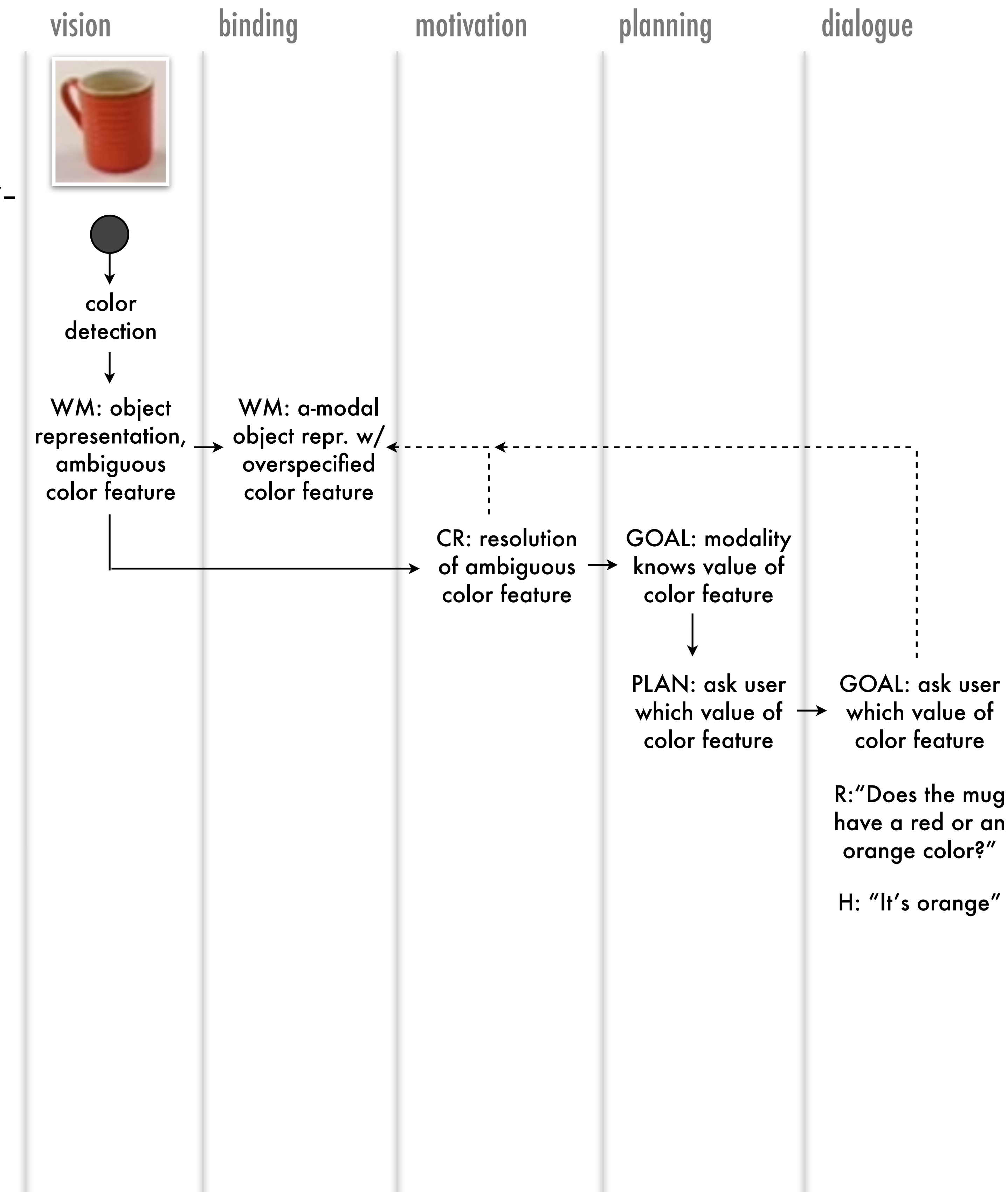
VISUAL LEARNING

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify



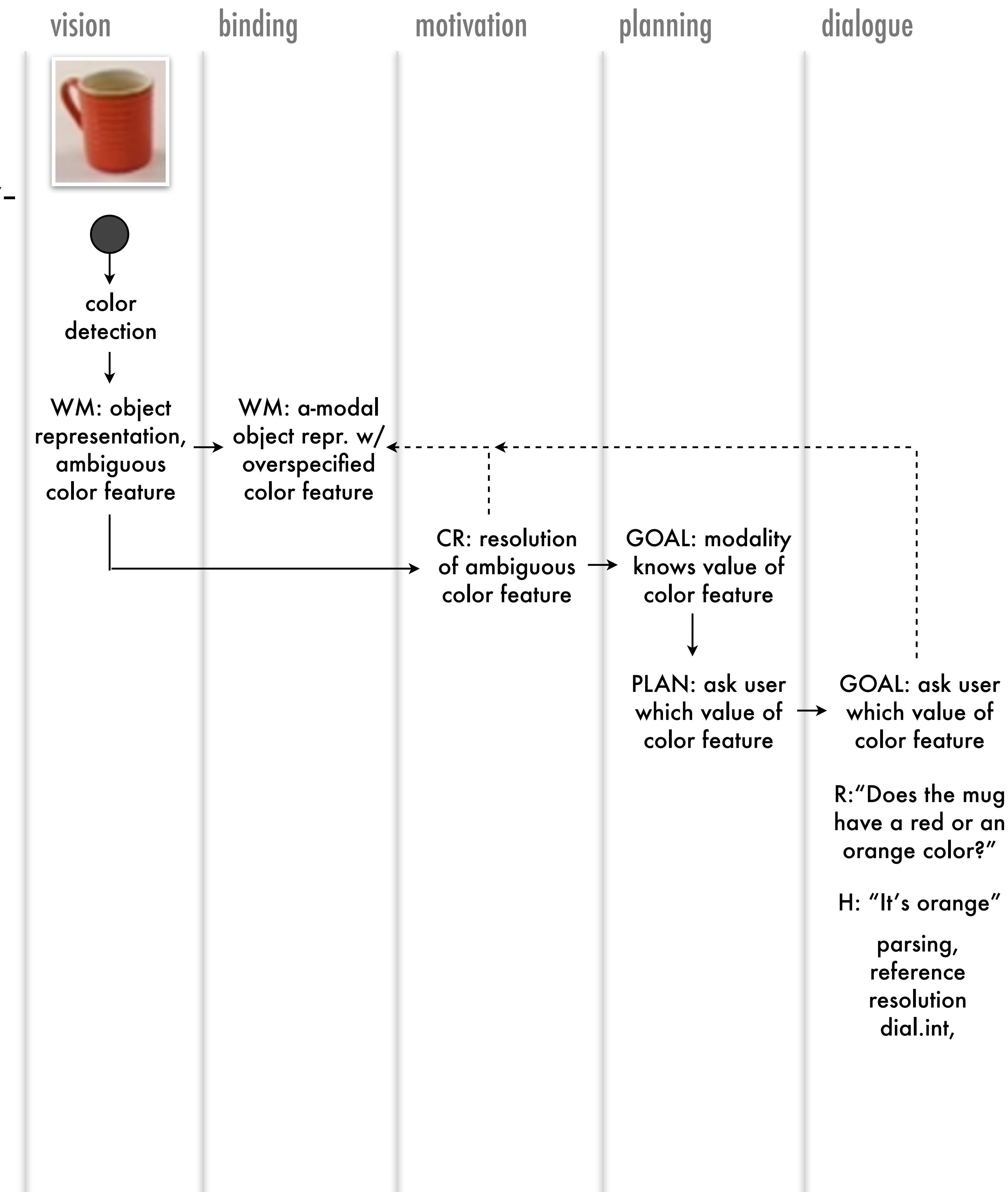
VISUAL LEARNING

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify

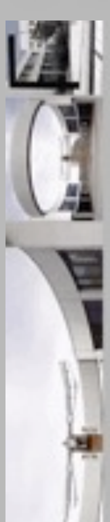


VISUAL LEARNING

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify

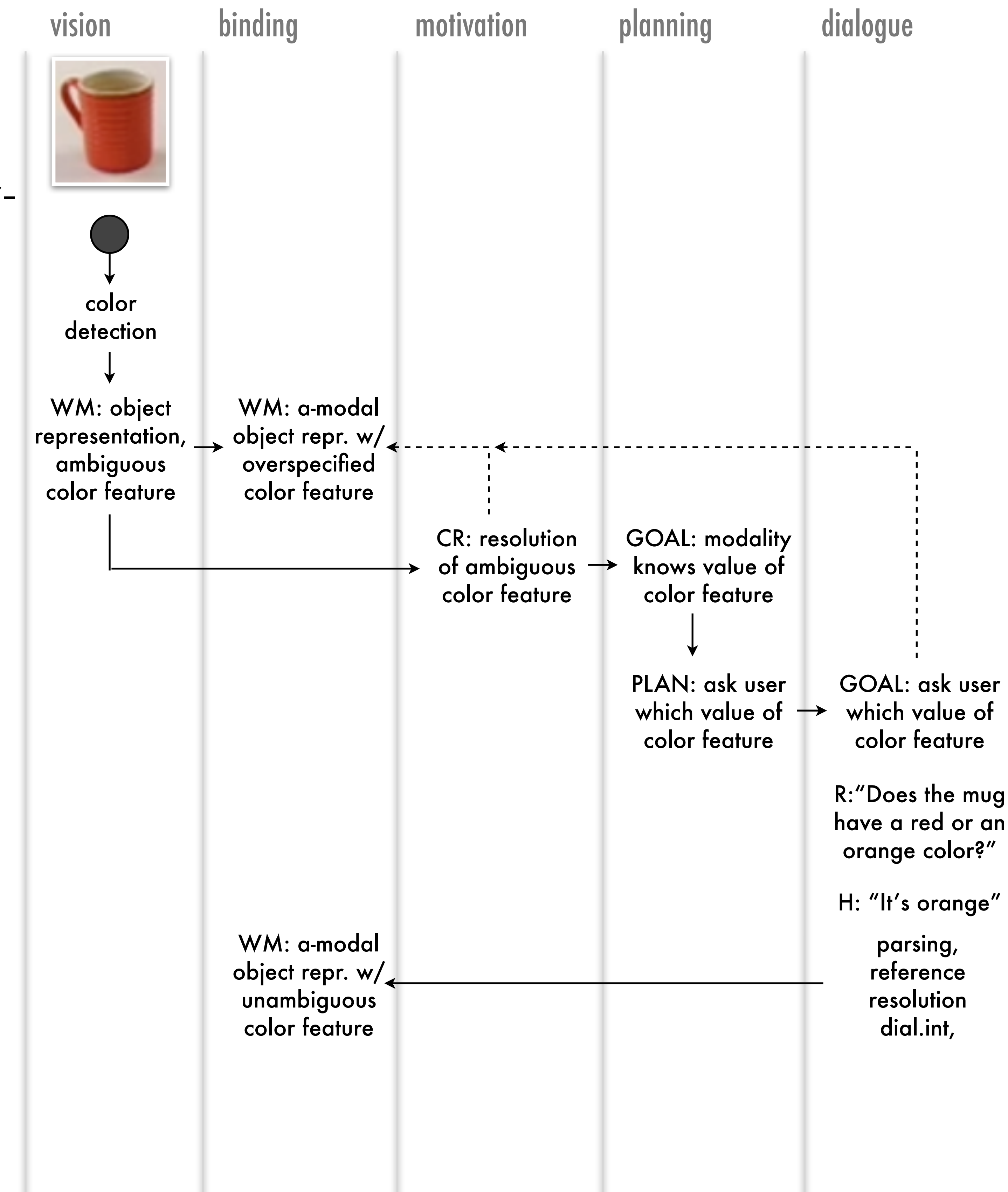


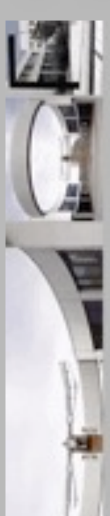
TALKING ROBOTS
LANGUAGE TECHNOLOGY, DFKI



VISUAL LEARNING

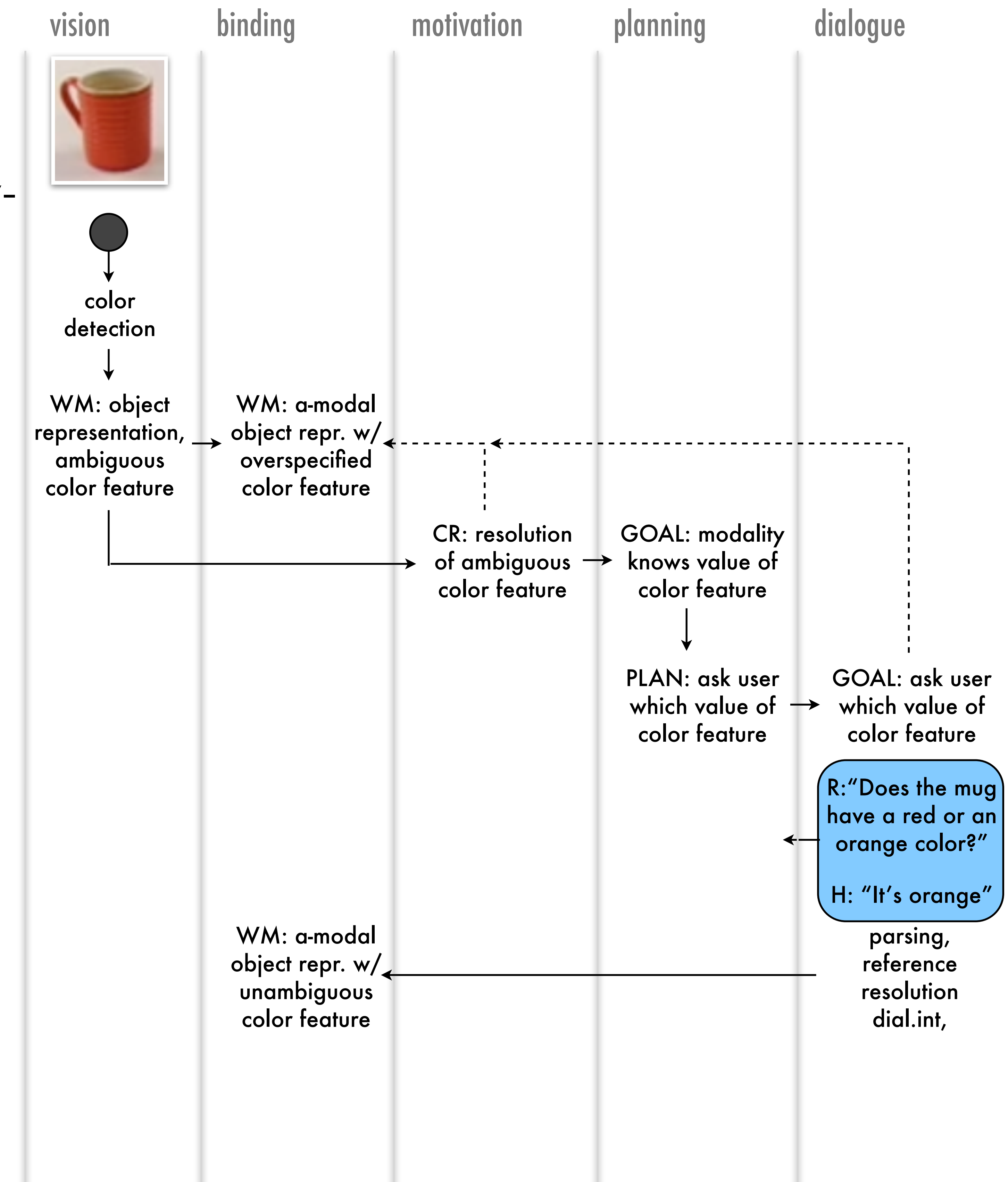
- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify



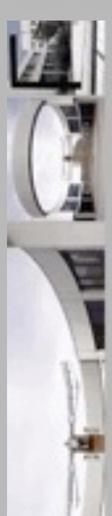


VISUAL LEARNING

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify

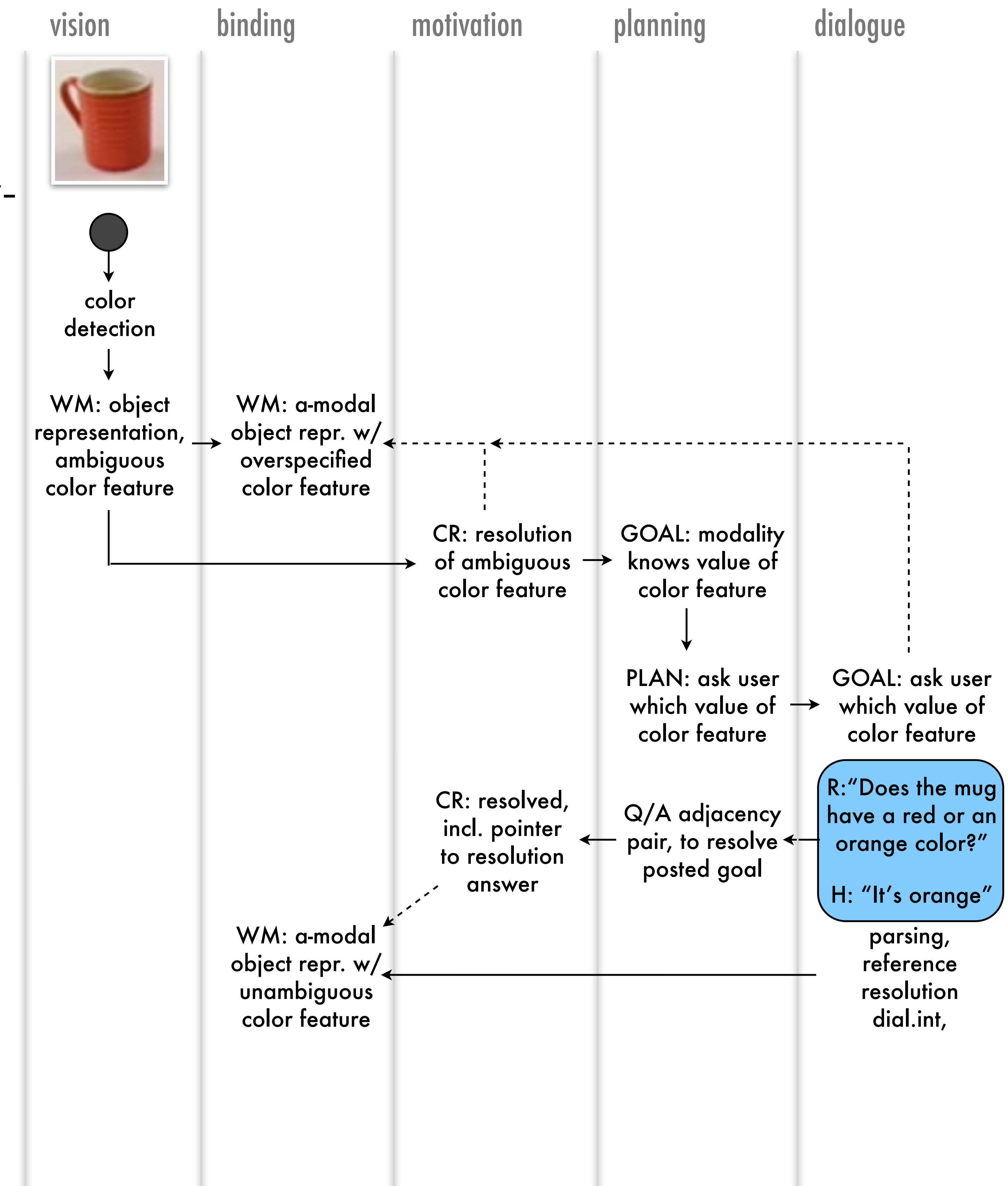


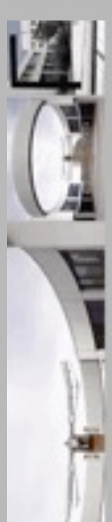
Talking Robots
Language Technology, DFKI



VISUAL LEARNING

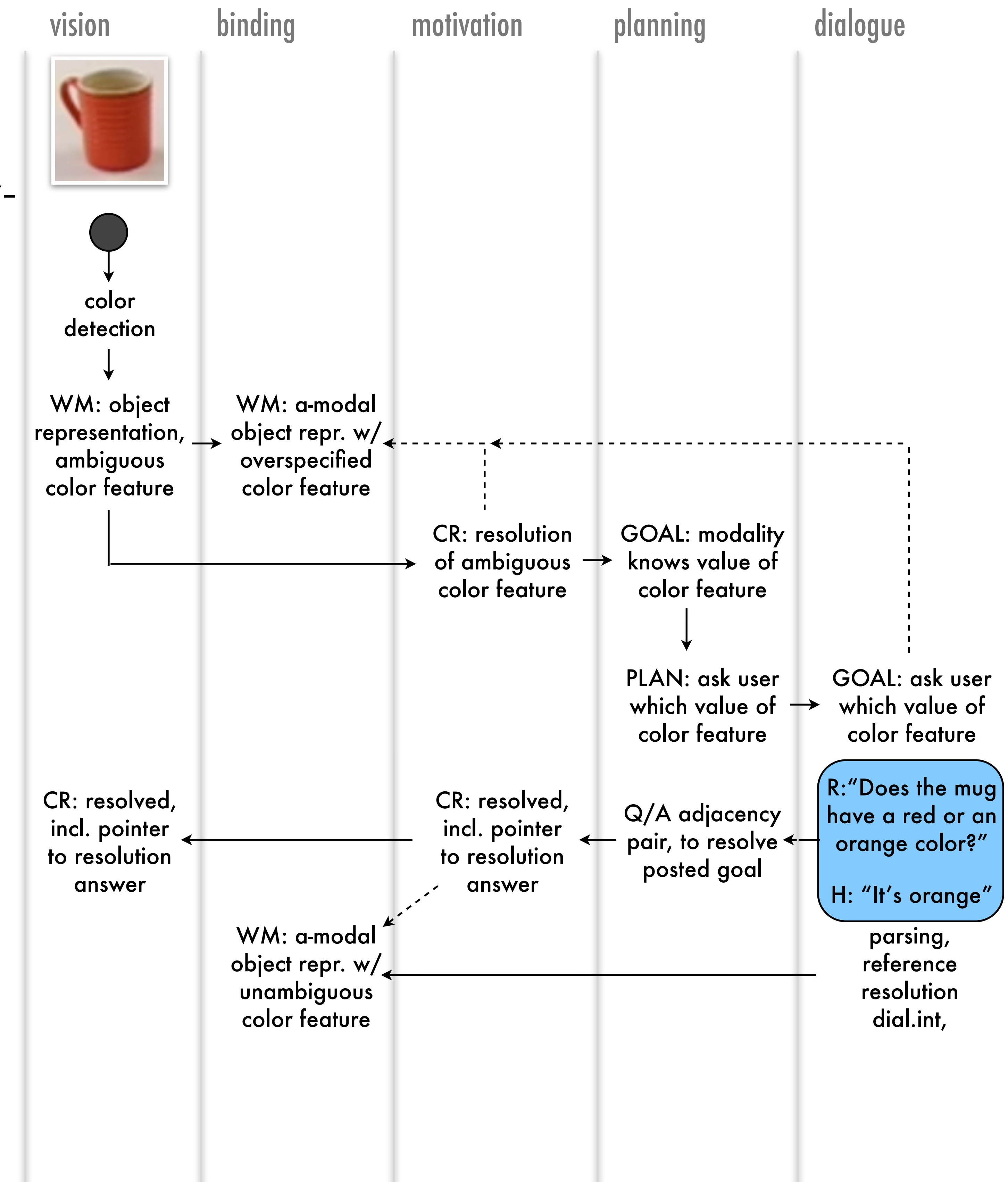
- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify





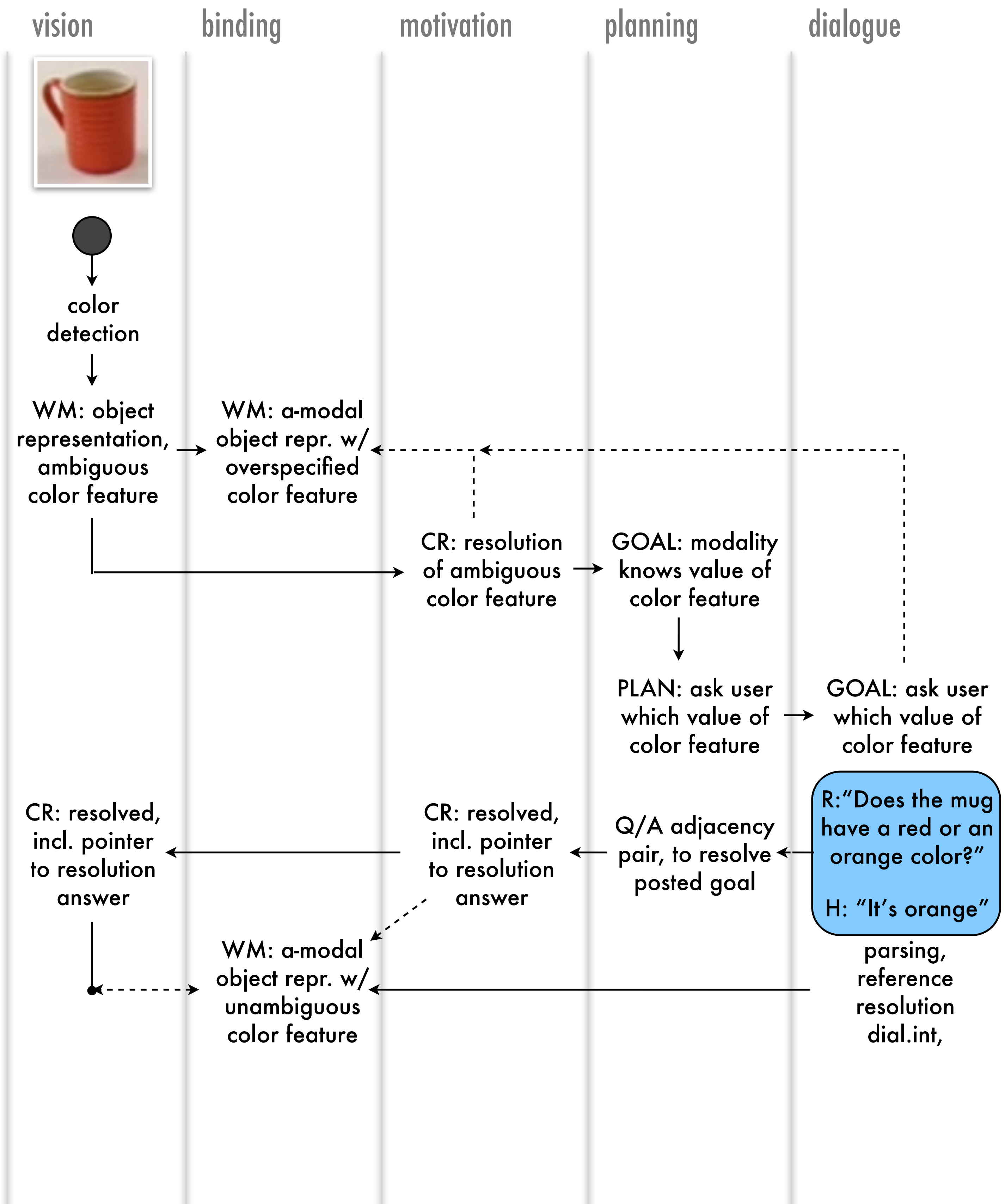
VISUAL LEARNING

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify



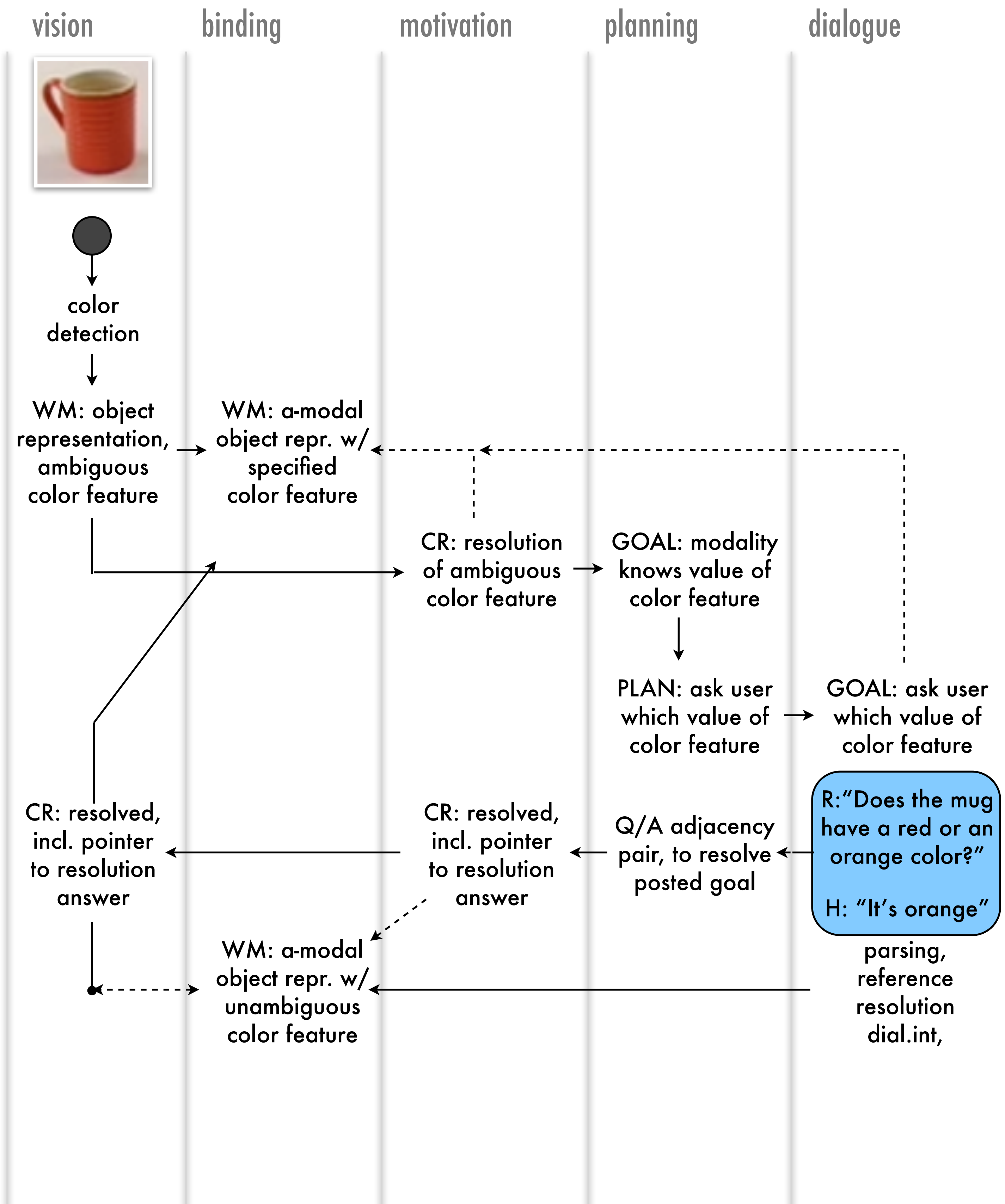
VISUAL LEARNING

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify



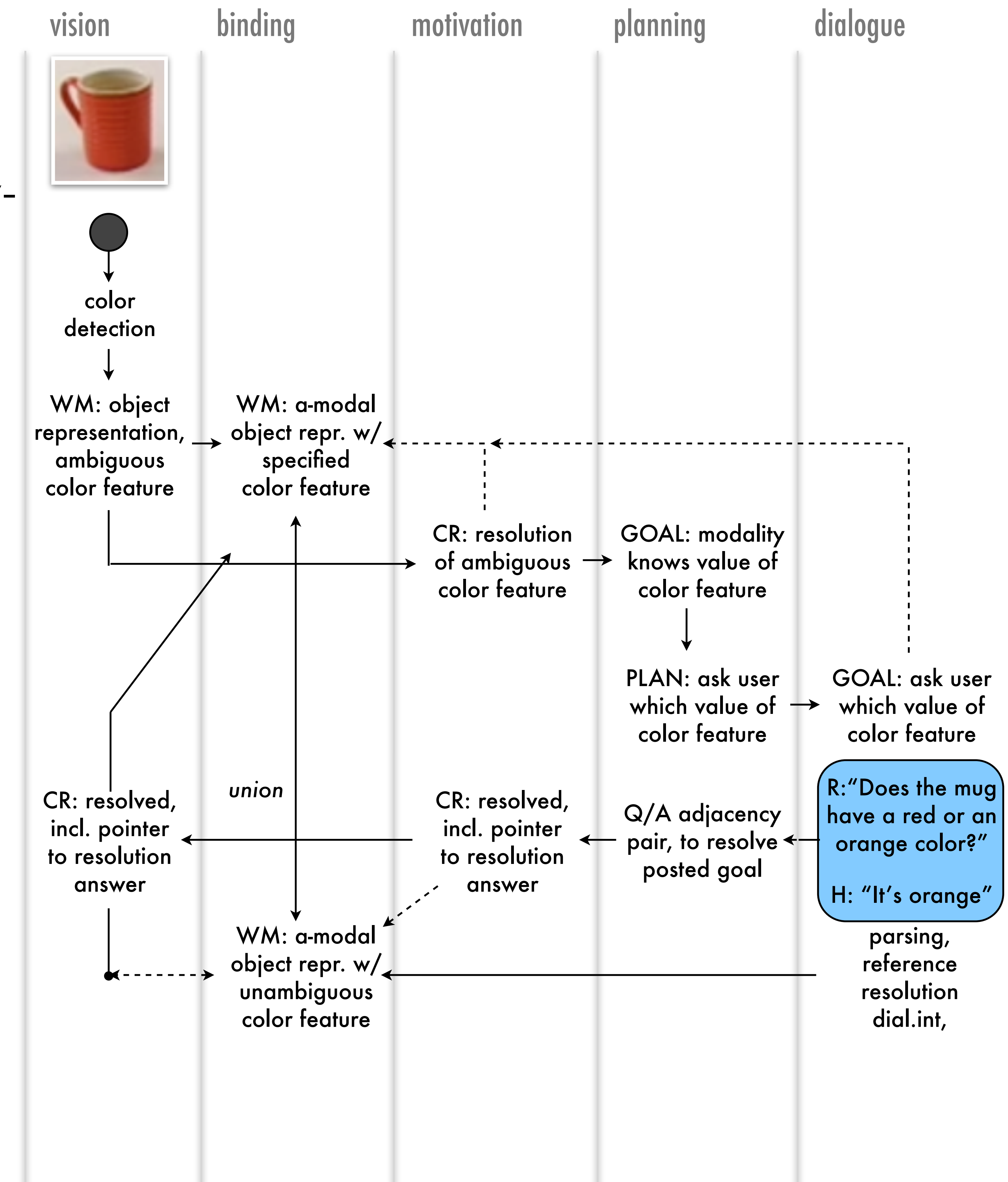
VISUAL LEARNING

- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify



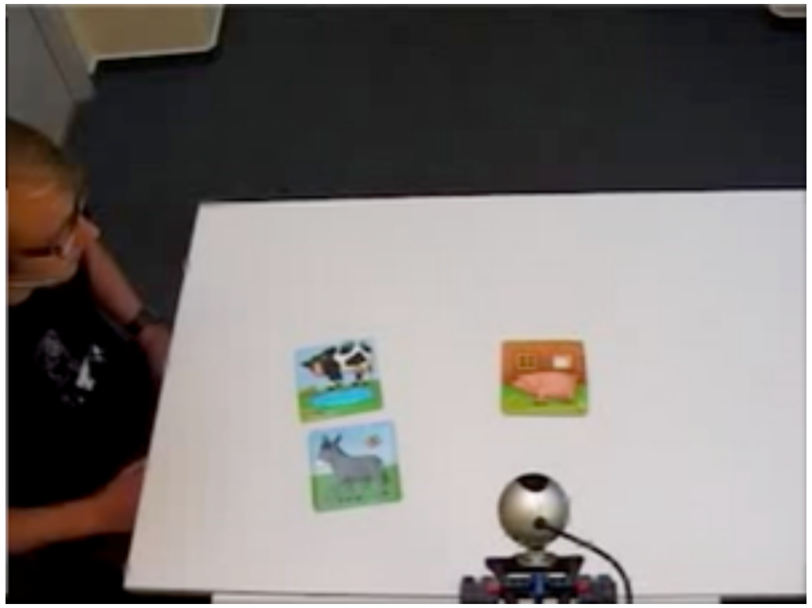
VISUAL LEARNING

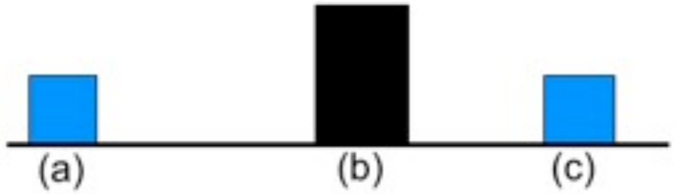
- Continual visual learning
  - Acquiring mappings between low-level features, and categorical descriptors
  - Continual, (inter-)active learning
  - Variation in forms of supervision: from unsupervised, over robot-initiated, to fully tutor-driven
- Use of clarification
  - Clarification of uncertainty in observations of training samples
  - Clarification requested at motivation level
  - Use of dialogue in interactive learning to clarify



Example: Spatial language

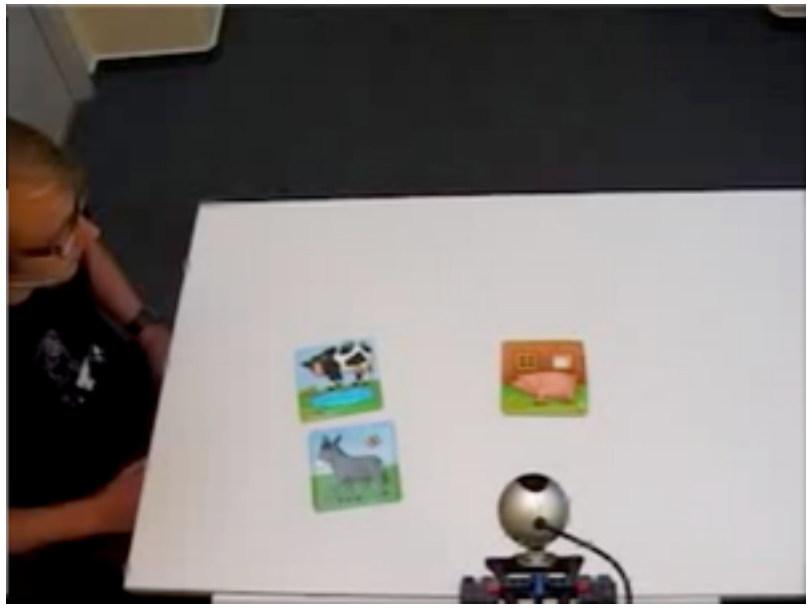
- Problem
  - “Uncertainty” as in quantitative variability in how something can be interpreted
  - Interplay between resolving referents, and resolving relations between referents
- Deeper issues
  - Quantitative (continuous) models of local scene structure
  - Complex interaction between material properties, “saliency,” task context, ...
- Development
  - Approaches covering production and comprehension of topological relations,
  - learnability of projective spatial relations

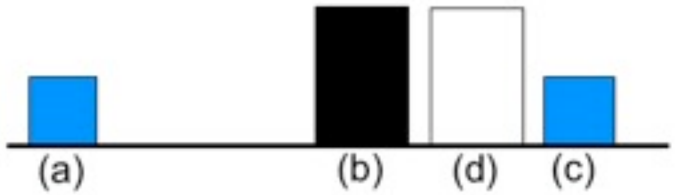




Example: Spatial language

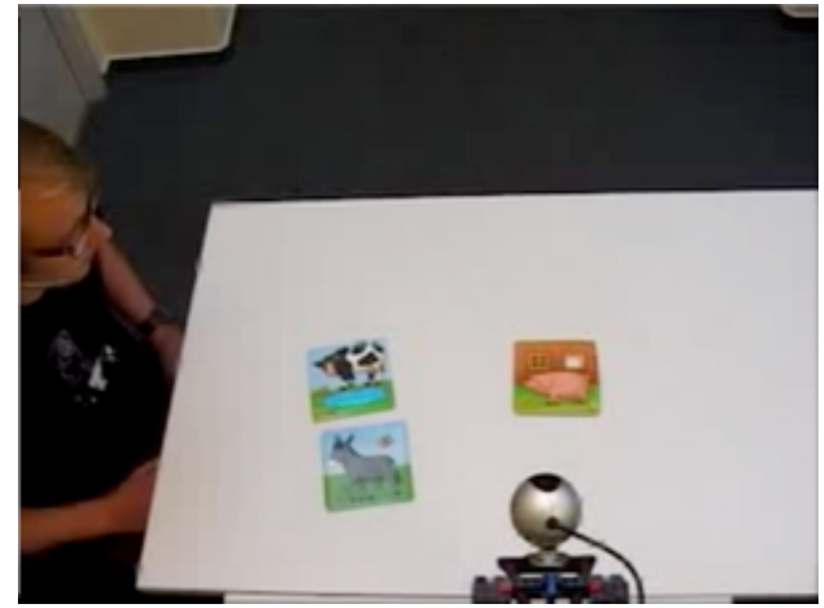
- Problem
  - "Uncertainty" as in quantitative variability in how something can be interpreted
  - Interplay between resolving referents, and resolving relations between referents
- Deeper issues
  - Quantitative (continuous) models of local scene structure
  - Complex interaction between material properties, "saliency," task context, ...
- Development
  - Approaches covering production and comprehension of topological relations,
  - learnability of projective spatial relations

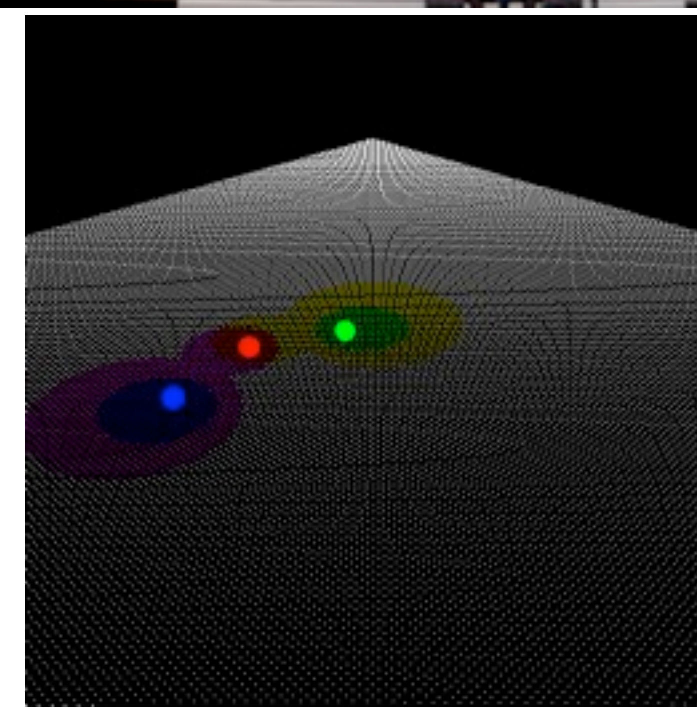




Example: Spatial language

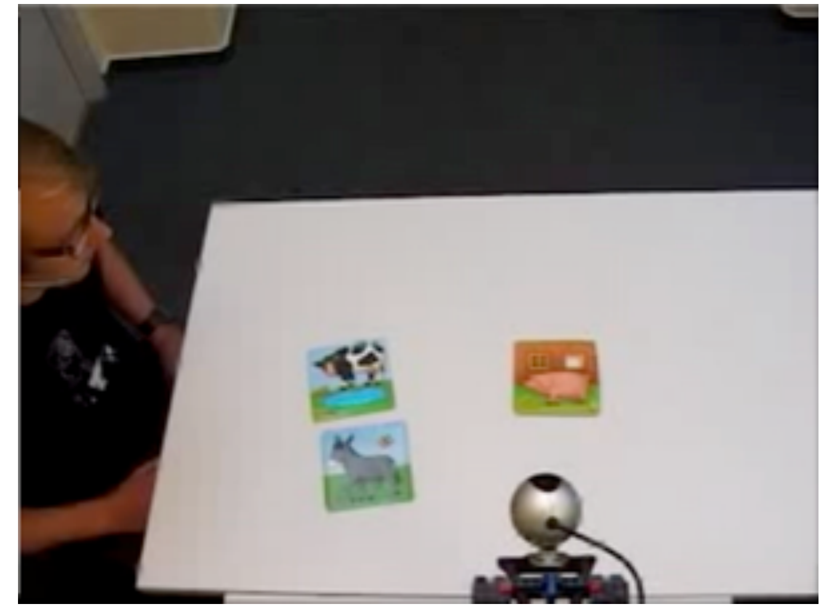
- Problem
  - "Uncertainty" as in quantitative variability in how something can be interpreted
  - Interplay between resolving referents, and resolving relations between referents
- Deeper issues
  - Quantitative (continuous) models of local scene structure
  - Complex interaction between material properties, "saliency," task context, ...
- Development
  - Approaches covering production and comprehension of topological relations,
  - learnability of projective spatial relations

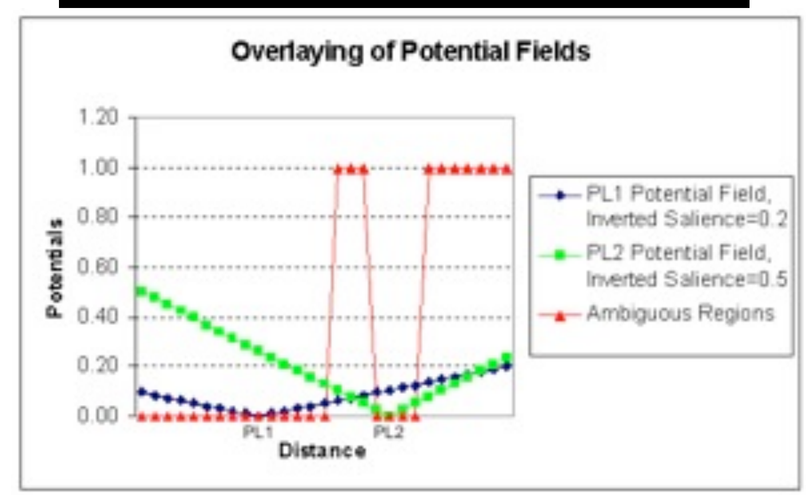


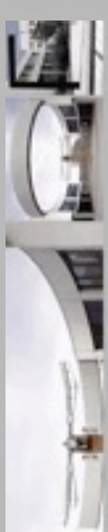


Example: Spatial language

- Problem
  - “Uncertainty” as in quantitative variability in how something can be interpreted
  - Interplay between resolving referents, and resolving relations between referents
- Deeper issues
  - Quantitative (continuous) models of local scene structure
  - Complex interaction between material properties, “saliency,” task context, ...
- Development
  - Approaches covering production and comprehension of topological relations,
  - learnability of projective spatial relations







Example: Spatial language

- Problem
  - "Uncertainty" as in quantitative variability in how something can be interpreted
  - Interplay between resolving referents, and resolving relations between referents
- Deeper issues
  - Quantitative (continuous) models of local scene structure
  - Complex interaction between material properties, "saliency," task context, ...
- Development
  - Approaches covering production and comprehension of topological relations,
  - learnability of projective spatial relations

COLING • ACL 2006

Proximity in Context: an empirically grounded computational model of proximity for processing topological spatial expressions*

John D. Kelleher
Dublin Institute of Technology
Dublin, Ireland
john.kelleher@comp.dit.ie

Geert-Jan M. Kruijff
DFKI GmbH
Saarbrücken, Germany
gj@dfki.de

Fintan J. Costello
University College Dublin
Dublin, Ireland
fintan.costello@ucd.ie

Abstract

The paper presents a new model for context-dependent interpretation of linguistic expressions about spatial proximity between objects in a natural scene. The paper discusses novel psycholinguistic experimental data that tests and verifies the model. The model has been implemented, and enables a conversational robot to identify objects in a scene through topological spatial relations (e.g. "X near Y"). The model can help motivate the choice between topological and projective prepositions.

1 Introduction

Our long-term goal is to develop conversational robots with which we can have natural, fluent situated dialog. An inherent aspect of such situated dialog is reference to aspects of the physical environment in which the agents are situated. In this paper, we present a computational model which provides a context-dependent analysis of the environment in terms of *spatial proximity*. We show how we can use this model to ground spatial language that uses topological prepositions ("the ball near the box") to identify objects in a scene.

Proximity is ubiquitous in situated dialog, but there are deeper "cognitive" reasons for why we need a context-dependent model of proximity to facilitate fluent dialog with a conversational robot. This has to do with the cognitive load that processing proximity expressions imposes. Consider the examples in (1). Psycholinguistic data indicates that a spatial proximity expression (1b) presents a heavier cognitive load than a referring expression identifying an object purely on physical features (1a) yet is easier to process than a projective expression (1c) (van der Sluis and Krahmer, 2004).

(1) a. the blue ball
b. the ball near the box
c. the ball to the right of the box

One explanation for this preference is that feature-based descriptions are easier to resolve perceptually, with a further distinction among features as given in Figure 1, cf. (Dale and Reiter, 1995). On the other hand, the interpretation and realization of spatial expressions requires effort and attention (Logan, 1994; Logan, 1995).

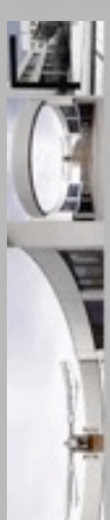
Similarly we can distinguish between the cognitive loads of processing different forms of spatial relations. Focusing on static prepositions, topological prepositions have a lower cognitive load than projective prepositions. Topological prepositions (e.g. "at", "near") describe proximity to an object. Projective prepositions (e.g. "above") describe a region in a particular direction from the object. Projective prepositions impose a higher cognitive load because we need to consider different spatial frames of reference (Krahmer and Theune, 1999; Moratz and Tenbrink, 2006). Now, if we want a robot to interact with other agents in a way that obeys the Principle of Minimal Cooperative Effort (Clark and Wilkes-Gibbs, 1986), it should adopt the simplest means to (spatially) refer to an object. However, research on spatial language in human-robot interaction has primarily focused on the use of projective prepositions.

Cognitive Load
HIGH
PROJECTIVE LOCATION
TOPOLOGICAL LOCATION
RELATIVE ADJECTIVE
ABSOLUTE ADJECTIVE
TYPE
LOW

Figure 1: Cognitive load

We currently lack a comprehensive model for topological prepositions. Without such a model,

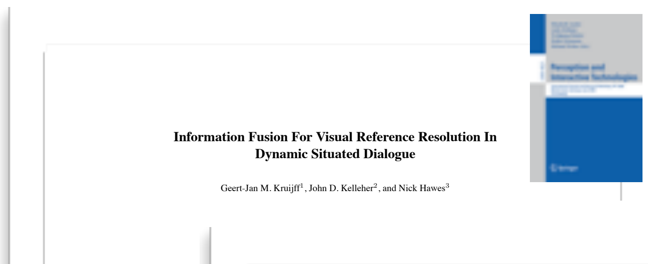
*The research reported here was supported by the CoSy project, EU FP6 IST "Cognitive Systems" FP6-004250-IP.



"Incompleteness"

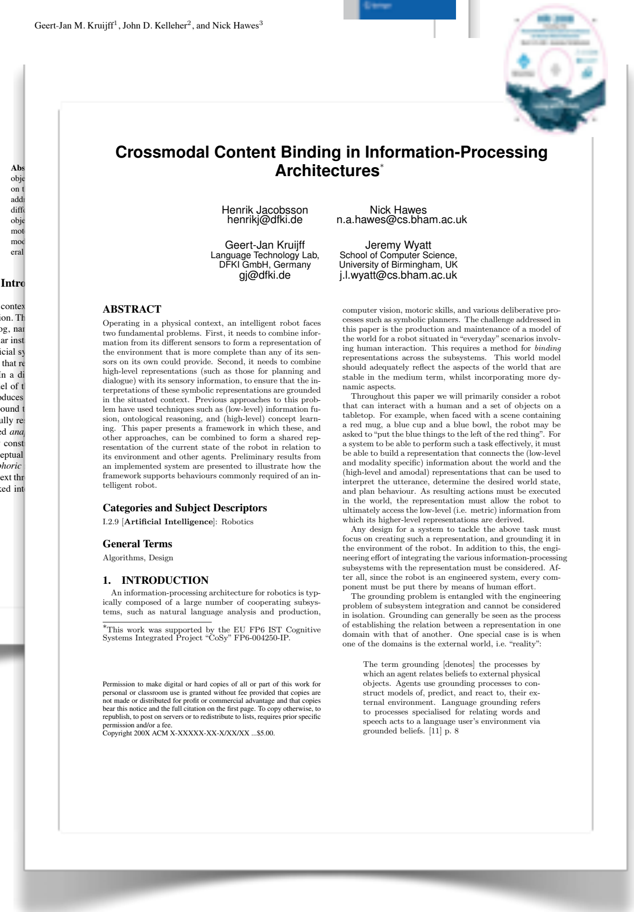
Mediation

- Ontology-based mediation
- Objects, features, uncertainty
- Reasoning with incompleteness



Information Fusion For Visual Reference Resolution In Dynamic Situated Dialogue

Geert-Jan M. Kruijff[1], John D. Kelleher[2], and Nick Hawes[3]



Crossmodal Content Binding in Information-Processing Architectures*

Henrik Jacobsson
henrikj@dfki.de

Nick Hawes
n.a.hawes@cs.bham.ac.uk

Geert-Jan Kruijff
Language Technology Lab,
DFKI GmbH, Germany
gj@dfki.de

Jeremy Wyatt
School of Computer Science,
University of Birmingham, UK
j.l.wyatt@cs.bham.ac.uk

ABSTRACT

Operating in a physical context, an intelligent robot faces two fundamental problems. First, it needs to combine information from its different sensors to form a representation of the environment that is more complete than any of its sensors on its own could provide. Second, it needs to combine high-level representations (such as those for planning and dialogue) with its sensory information, to ensure that the interpretations of these symbolic representations are grounded in the situated context. Previous approaches to this problem have used techniques such as (low-level) information fusion, ontological reasoning, and (high-level) concept learning. This paper presents a framework in which these, and other approaches, can be combined to form a shared representation of the current state of the robot in relation to its environment and other agents. Preliminary results from an implemented system are presented to illustrate how the framework supports behaviours commonly required of an intelligent robot.

Categories and Subject Descriptors

I.2.9 [**Artificial Intelligence**]: Robotics

General Terms

Algorithms, Design

1. INTRODUCTION

An information-processing architecture for robotics is typically composed of a large number of cooperating subsystems, such as natural language analysis and production, computer vision, motoric skills, and various deliberative processes such as symbolic planners. The challenge addressed in this paper is the production and maintenance of a model of the world for a robot situated in "everyday" scenarios involving human interaction. This requires a method for *binding* representations across the subsystems. This world model should adequately reflect the aspects of the world that are stable in the medium term, whilst incorporating more dynamic aspects.

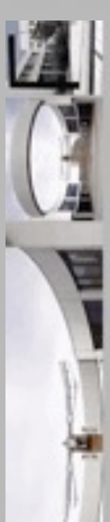
Throughout this paper we will primarily consider a robot that can interact with a human and a set of objects on a tabletop. For example, when faced with a scene containing a red mug, a blue cup and a blue bowl, the robot may be asked to "put the blue things to the left of the red thing". For a system to be able to perform such a task effectively, it must be able to build a representation that connects the (low-level and modality specific) information about the world and the (high-level and amodal) representations that can be used to interpret the utterance, determine the desired world state, and plan behaviour. As resulting actions must be executed in the world, the representation must allow the robot to ultimately access the low-level (i.e. metric) information from which its higher-level representations are derived.

Any design for a system to tackle the above task must focus on creating such a representation, and grounding it in the environment of the robot. In addition to this, the engineering effort of integrating the various information-processing subsystems with the representation must be considered. After all, since the robot is an engineered system, every component must be put there by means of human effort.

The grounding problem is entangled with the engineering problem of subsystem integration and cannot be considered in isolation. Grounding can generally be seen as the process of establishing the relation between a representation in one domain with that of another. One special case is is when one of the domains is the external world, i.e. "reality":

> The term grounding [denotes] the processes by which an agent relates beliefs to external physical objects. Agents use grounding processes to construct models of, predict, and react to, their external environment. Language grounding refers to processes specialised for relating words and speech acts to a language user's environment via grounded beliefs. [11] p. 8

*This work was supported by the EU FP6 IST Cognitive Systems Integrated Project "CoSy" FP6-004250-IP.



Mediation

- Ontology-based mediation
- Objects, features, uncertainty
- Reasoning with incompleteness

Self-Understanding & Self-Extension: A Systems and Representational Approach

Jeremy L. Wyatt, Alper Aydemir, Michael Brenner, Marc Hanheide, Nick Hawes, Patric Jensfelt, Matej Kristan, Geert-Jan M. Kruijff, Pierre Lison, Andrzej Pronobis, Kristoffer Sjöö, Alen Vrečko, Hendrik Zender, Michael Zillich, Danijel Skočaj

***Abstract*—There are many different approaches to building a system that can engage in autonomous mental development. In this paper we present an approach based on what we term *self-understanding*, by which we mean the explicit representation of and reasoning about what a system does and doesn't know, and how that knowledge changes under action. We present an architecture and a set of representations used in two robot systems that exhibit a limited degree of autonomous mental development, which we term *self-extension*. The contributions include: representations of gaps and uncertainty for specific kinds of knowledge, and a goal management and planning system for setting and achieving learning goals.**

***Index Terms*—robotics, robot learning, architectures, representations**

I. Introduction

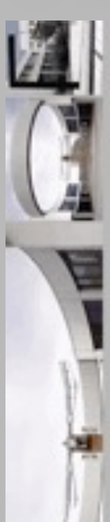
WHAT is needed for an agent to learn in a truly autonomous fashion? Autonomous learning requires that the agent pick its own learning goals. One way to achieve this is to give that agent representations of what it knows and doesn't know, and to make it reason with these representations to set its own *epistemic goals*. An epistemic goal is a goal to be in a certain knowledge state. This paper describes this approach to autonomous mental development. We present an architecture, together with a set of representations that explicitly capture what the robot and other agents do and don't know at any time, i.e. representations of their epistemic state. We also describe representations of how this epistemic state will change under action. Such representations, together with algorithms for reasoning about them confer a degree of *self-understanding*, and allow the agent to plan how to extend its abilities, or knowledge of the environment, i.e. *self-extension*. We also describe a goal management system that allows the robot to choose quickly between different epistemic goals. This mechanism is necessary to allow our approach to scale, since if a robot generates many possible learning goals the time taken to plan for them all will be too great.

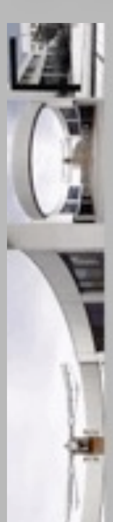
We first define self-understanding and self-extension as we see them. To do this it is necessary to characterise the different types of incompleteness in knowledge that will be represented. We use *incompleteness* as an umbrella term to cover many different types of *knowledge gaps* and *uncertainty about knowledge*. We can construct a typology of incompleteness in knowledge based on three dimensions of variability. These are *the nature of the incompleteness*, the *type of knowledge that is incomplete*, and whether the incompleteness is represented in a *quantitative or qualitative* manner.

With regard to the nature of the incompleteness, in the simplest case we may have a variable or variables that are part of a model of the world and which have a defined set of possible values or hypotheses from which the true value is known to be drawn. We refer to this as *simple uncertainty*. We can also have *uncertainty about the number of variables* needed in a model, i.e. about the model complexity. Finally we can also have cases where the agent knows that a variable is of an unexperienced class, i.e. there is *novelty*. This can include cases where the variables are continuous but where the observation models for a class are quite confident and do not generalise well to some new observation. The type of knowledge that is incomplete may vary enormously. Five simple types that cover a variety of cases include contingent knowledge about the current world *state*, *structural* knowledge about the universal relationships between variables, *procedural* knowledge about how to act in certain situations to achieve certain goals, knowledge consisting of *predictions* of action outcomes or events, and knowledge about their *causes*. Finally there is a question about whether the representation is qualitative or quantitative. In qualitative representations we simply have a set of possible values for the variable, or a statement that the variable value is unknown, or knowledge that there may be many variables that are unmodelled. In quantitative representations we will have some kind of scalar values attached to hypotheses (such as whether there is novelty or not), and in our case these will typically be probabilities. Note that by a *quantitative gap* or *quantitative uncertainty* we do not mean that the underlying space for the variable is continuous or discrete, but instead that the way the incompleteness is represented involves an expression of a degree of preference for one hypothesis or statement versus another.

Given this brief characterisation of different types of incompleteness we can define self-understanding and self-extension compactly as follows. A system with *self-understanding* is any system with explicit representations that captures some

Jeremy L. Wyatt, Marc Hanheide and Nick Hawes are with the University of Birmingham, email: {jlw,nah,m.hanheide}@cs.bham.ac.uk
Michael Brenner is with Albert-Ludwigs-Universität, email: brenner@informatik.uni-freiburg.de
Pierre Lison, Geert-Jan M. Kruijff and Hendrik Zender are with DFKI, Saarbrücken, Germany, email: {plison,gj,zender}@dfki.de
Patric Jensfelt, Andrzej Pronobis, Kristoffer Sjöö and Alper Aydemir are with KTH Stockholm, email: {patric,krsj,pronobis,aydemir}@csc.kth.se
Matej Kristan, Alen Vrečko and Danijel Skočaj are with University of Ljubljana, email: {matej.kristan,alen.vrecko,danijel.skocaj}@fri.uni-lj.si
Michael Zillich is with Vienna University of Technology, email: zillich@acin.tuwien.ac.at
Manuscript received February 28, 2010, revised Sept 13, 2010





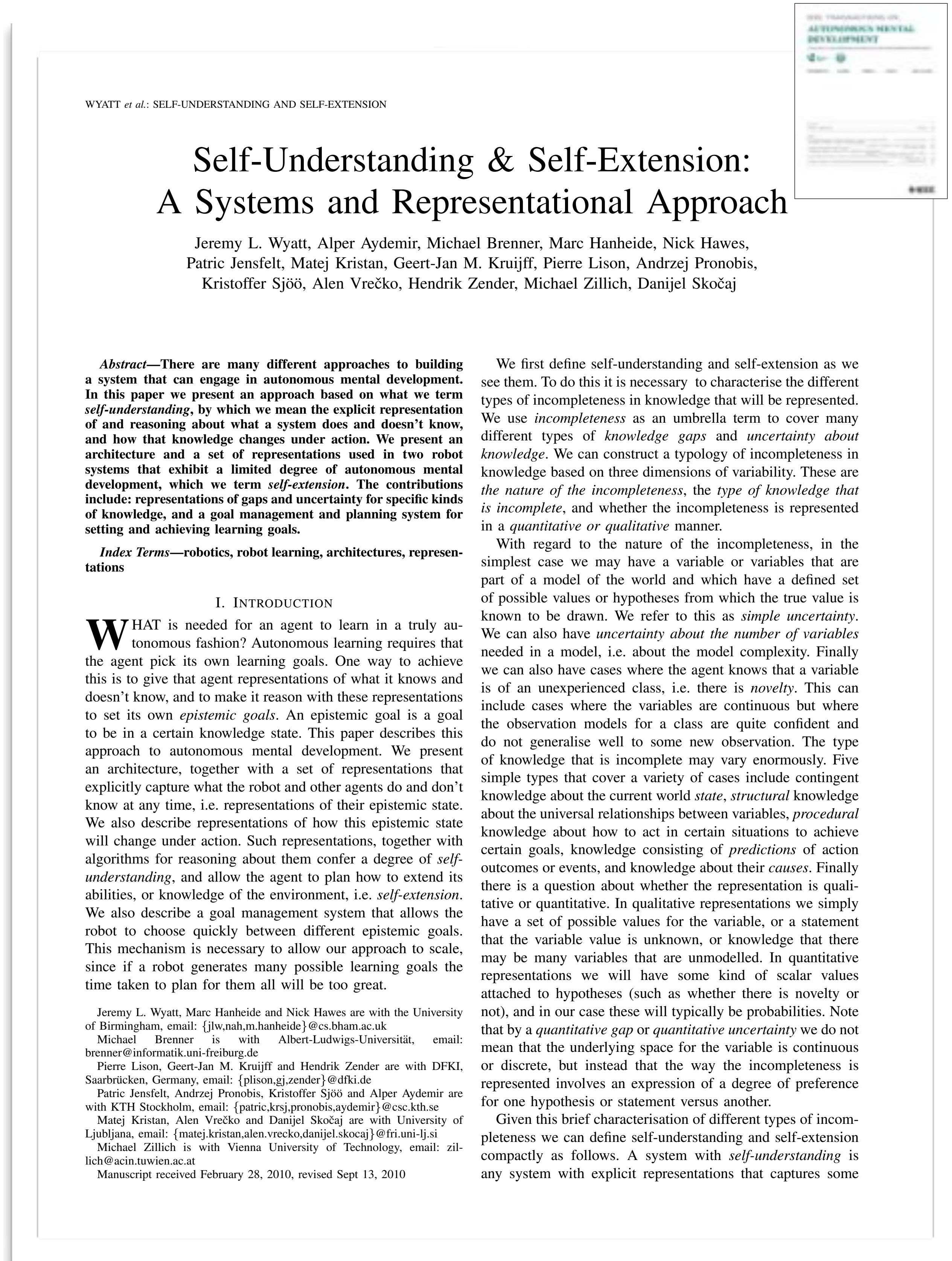
Mediation

- Ontology-based mediation
- Objects, features, uncertainty
- Reasoning with incompleteness

Conceptual meaning:
concept(**box**) & ...

Modal meaning:
c1(**box**) & f1(...) ..

Mapping



Self-Understanding & Self-Extension: A Systems and Representational Approach

Jeremy L. Wyatt, Alper Aydemir, Michael Brenner, Marc Hanheide, Nick Hawes,
Patric Jensfelt, Matej Kristan, Geert-Jan M. Kruijff, Pierre Lison, Andrzej Pronobis,
Kristoffer Sjöö, Alen Vrečko, Hendrik Zender, Michael Zillich, Danijel Skočaj

***Abstract*—There are many different approaches to building a system that can engage in autonomous mental development. In this paper we present an approach based on what we term *self-understanding*, by which we mean the explicit representation of and reasoning about what a system does and doesn't know, and how that knowledge changes under action. We present an architecture and a set of representations used in two robot systems that exhibit a limited degree of autonomous mental development, which we term *self-extension*. The contributions include: representations of gaps and uncertainty for specific kinds of knowledge, and a goal management and planning system for setting and achieving learning goals.**

***Index Terms*—robotics, robot learning, architectures, representations**

I. Introduction

WHAT is needed for an agent to learn in a truly autonomous fashion? Autonomous learning requires that the agent pick its own learning goals. One way to achieve this is to give that agent representations of what it knows and doesn't know, and to make it reason with these representations to set its own *epistemic goals*. An epistemic goal is a goal to be in a certain knowledge state. This paper describes this approach to autonomous mental development. We present an architecture, together with a set of representations that explicitly capture what the robot and other agents do and don't know at any time, i.e. representations of their epistemic state. We also describe representations of how this epistemic state will change under action. Such representations, together with algorithms for reasoning about them confer a degree of *self-understanding*, and allow the agent to plan how to extend its abilities, or knowledge of the environment, i.e. *self-extension*. We also describe a goal management system that allows the robot to choose quickly between different epistemic goals. This mechanism is necessary to allow our approach to scale, since if a robot generates many possible learning goals the time taken to plan for them all will be too great.

Jeremy L. Wyatt, Marc Hanheide and Nick Hawes are with the University of Birmingham, email: {jlw,nah,m.hanheide}@cs.bham.ac.uk
Michael Brenner is with Albert-Ludwigs-Universität, email: brenner@informatik.uni-freiburg.de
Pierre Lison, Geert-Jan M. Kruijff and Hendrik Zender are with DFKI, Saarbrücken, Germany, email: {plison,gj,zender}@dfki.de
Patric Jensfelt, Andrzej Pronobis, Kristoffer Sjöö and Alper Aydemir are with KTH Stockholm, email: {patric,krsj,pronobis,aydemir}@csc.kth.se
Matej Kristan, Alen Vrečko and Danijel Skočaj are with University of Ljubljana, email: {matej.kristan,alen.vrecko,danijel.skocaj}@fri.uni-lj.si
Michael Zillich is with Vienna University of Technology, email: zillich@acin.tuwien.ac.at
Manuscript received February 28, 2010, revised Sept 13, 2010

We first define self-understanding and self-extension as we see them. To do this it is necessary to characterise the different types of incompleteness in knowledge that will be represented. We use *incompleteness* as an umbrella term to cover many different types of *knowledge gaps* and *uncertainty about knowledge*. We can construct a typology of incompleteness in knowledge based on three dimensions of variability. These are *the nature of the incompleteness*, the *type of knowledge that is incomplete*, and whether the incompleteness is represented in a *quantitative or qualitative* manner.

With regard to the nature of the incompleteness, in the simplest case we may have a variable or variables that are part of a model of the world and which have a defined set of possible values or hypotheses from which the true value is known to be drawn. We refer to this as *simple uncertainty*. We can also have *uncertainty about the number of variables* needed in a model, i.e. about the model complexity. Finally we can also have cases where the agent knows that a variable is of an unexperienced class, i.e. there is *novelty*. This can include cases where the variables are continuous but where the observation models for a class are quite confident and do not generalise well to some new observation. The type of knowledge that is incomplete may vary enormously. Five simple types that cover a variety of cases include contingent knowledge about the current world *state*, *structural* knowledge about the universal relationships between variables, *procedural* knowledge about how to act in certain situations to achieve certain goals, knowledge consisting of *predictions* of action outcomes or events, and knowledge about their *causes*. Finally there is a question about whether the representation is qualitative or quantitative. In qualitative representations we simply have a set of possible values for the variable, or a statement that the variable value is unknown, or knowledge that there may be many variables that are unmodelled. In quantitative representations we will have some kind of scalar values attached to hypotheses (such as whether there is novelty or not), and in our case these will typically be probabilities. Note that by a *quantitative gap* or *quantitative uncertainty* we do not mean that the underlying space for the variable is continuous or discrete, but instead that the way the incompleteness is represented involves an expression of a degree of preference for one hypothesis or statement versus another.

Given this brief characterisation of different types of incompleteness we can define self-understanding and self-extension compactly as follows. A system with *self-understanding* is any system with explicit representations that captures some

Mediation

- Ontology-based mediation
- Objects, features, uncertainty
- Reasoning with incompleteness

Conceptual reasoning:
concept(**box**) & *color*(X)

↕

Conceptual meaning:
concept(**box**) & ...

Modal meaning:
c1(**box**) & f1(...) ..

Mapping

Self-Understanding & Self-Extension: A Systems and Representational Approach

Jeremy L. Wyatt, Alper Aydemir, Michael Brenner, Marc Hanheide, Nick Hawes, Patric Jensfelt, Matej Kristan, Geert-Jan M. Kruijff, Pierre Lison, Andrzej Pronobis, Kristoffer Sjöö, Alen Vrečko, Hendrik Zender, Michael Zillich, Danijel Skočaj

***Abstract*—There are many different approaches to building a system that can engage in autonomous mental development. In this paper we present an approach based on what we term *self-understanding*, by which we mean the explicit representation of and reasoning about what a system does and doesn't know, and how that knowledge changes under action. We present an architecture and a set of representations used in two robot systems that exhibit a limited degree of autonomous mental development, which we term *self-extension*. The contributions include: representations of gaps and uncertainty for specific kinds of knowledge, and a goal management and planning system for setting and achieving learning goals.**

***Index Terms*—robotics, robot learning, architectures, representations**

I. Introduction

WHAT is needed for an agent to learn in a truly autonomous fashion? Autonomous learning requires that the agent pick its own learning goals. One way to achieve this is to give that agent representations of what it knows and doesn't know, and to make it reason with these representations to set its own *epistemic goals*. An epistemic goal is a goal to be in a certain knowledge state. This paper describes this approach to autonomous mental development. We present an architecture, together with a set of representations that explicitly capture what the robot and other agents do and don't know at any time, i.e. representations of their epistemic state. We also describe representations of how this epistemic state will change under action. Such representations, together with algorithms for reasoning about them confer a degree of *self-understanding*, and allow the agent to plan how to extend its abilities, or knowledge of the environment, i.e. *self-extension*. We also describe a goal management system that allows the robot to choose quickly between different epistemic goals. This mechanism is necessary to allow our approach to scale, since if a robot generates many possible learning goals the time taken to plan for them all will be too great.

Jeremy L. Wyatt, Marc Hanheide and Nick Hawes are with the University of Birmingham, email: {jlw,nah,m.hanheide}@cs.bham.ac.uk
Michael Brenner is with Albert-Ludwigs-Universität, email: brenner@informatik.uni-freiburg.de
Pierre Lison, Geert-Jan M. Kruijff and Hendrik Zender are with DFKI, Saarbrücken, Germany, email: {plison,gj,zender}@dfki.de
Patric Jensfelt, Andrzej Pronobis, Kristoffer Sjöö and Alper Aydemir are with KTH Stockholm, email: {patric,krsj,pronobis,aydemir}@csc.kth.se
Matej Kristan, Alen Vrečko and Danijel Skočaj are with University of Ljubljana, email: {matej.kristan,alen.vrecko,danijel.skocaj}@fri.uni-lj.si
Michael Zillich is with Vienna University of Technology, email: zillich@acin.tuwien.ac.at
Manuscript received February 28, 2010, revised Sept 13, 2010

We first define self-understanding and self-extension as we see them. To do this it is necessary to characterise the different types of incompleteness in knowledge that will be represented. We use *incompleteness* as an umbrella term to cover many different types of *knowledge gaps* and *uncertainty about knowledge*. We can construct a typology of incompleteness in knowledge based on three dimensions of variability. These are *the nature of the incompleteness*, the *type of knowledge that is incomplete*, and whether the incompleteness is represented in a *quantitative or qualitative* manner.

With regard to the nature of the incompleteness, in the simplest case we may have a variable or variables that are part of a model of the world and which have a defined set of possible values or hypotheses from which the true value is known to be drawn. We refer to this as *simple uncertainty*. We can also have *uncertainty about the number of variables* needed in a model, i.e. about the model complexity. Finally we can also have cases where the agent knows that a variable is of an unexperienced class, i.e. there is *novelty*. This can include cases where the variables are continuous but where the observation models for a class are quite confident and do not generalise well to some new observation. The type of knowledge that is incomplete may vary enormously. Five simple types that cover a variety of cases include contingent knowledge about the current world *state*, *structural* knowledge about the universal relationships between variables, *procedural* knowledge about how to act in certain situations to achieve certain goals, knowledge consisting of *predictions* of action outcomes or events, and knowledge about their *causes*. Finally there is a question about whether the representation is qualitative or quantitative. In qualitative representations we simply have a set of possible values for the variable, or a statement that the variable value is unknown, or knowledge that there may be many variables that are unmodelled. In quantitative representations we will have some kind of scalar values attached to hypotheses (such as whether there is novelty or not), and in our case these will typically be probabilities. Note that by a *quantitative gap* or *quantitative uncertainty* we do not mean that the underlying space for the variable is continuous or discrete, but instead that the way the incompleteness is represented involves an expression of a degree of preference for one hypothesis or statement versus another.

Given this brief characterisation of different types of incompleteness we can define self-understanding and self-extension compactly as follows. A system with *self-understanding* is any system with explicit representations that captures some

Mediation

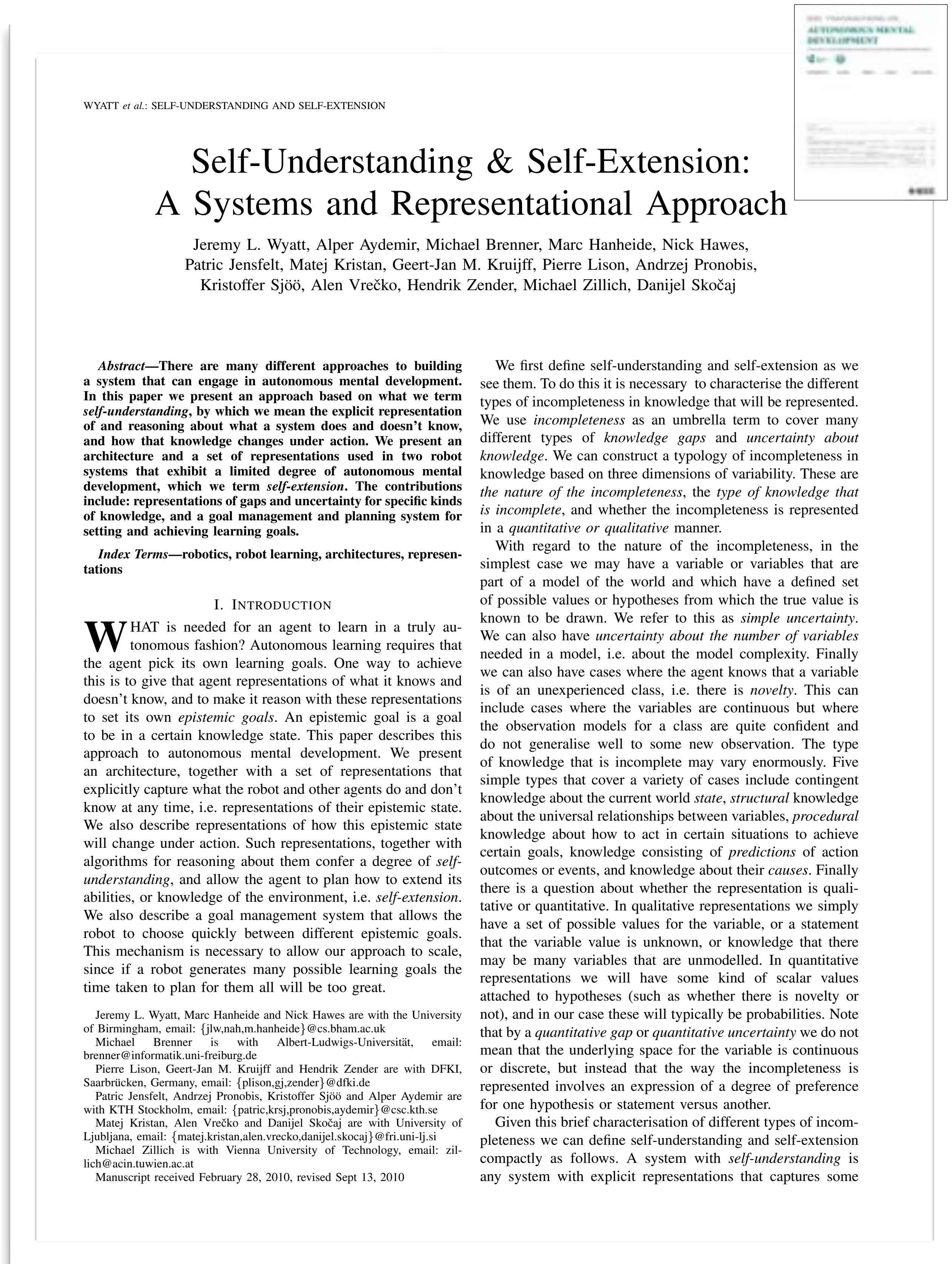
- Ontology-based mediation
- Objects, features, uncertainty
- Reasoning with incompleteness

Conceptual reasoning:
concept(**box**) & *color*(X)

Conceptual meaning:
concept(**box**) & ...

Modal meaning:
c1(**box**) & f1(...) ..

Mapping



WYATT *et al.*: SELF-UNDERSTANDING AND SELF-EXTENSION

Self-Understanding & Self-Extension: A Systems and Representational Approach

Jeremy L. Wyatt, Alper Aydemir, Michael Brenner, Marc Hanheide, Nick Hawes, Patric Jensfelt, Matej Kristan, Geert-Jan M. Kruijff, Pierre Lison, Andrzej Pronobis, Kristoffer Sjöö, Alen Vrečko, Hendrik Zender, Michael Zillich, Danijel Skočaj

***Abstract*—There are many different approaches to building a system that can engage in autonomous mental development. In this paper we present an approach based on what we term *self-understanding*, by which we mean the explicit representation of and reasoning about what a system does and doesn't know, and how that knowledge changes under action. We present an architecture and a set of representations used in two robot systems that exhibit a limited degree of autonomous mental development, which we term *self-extension*. The contributions include: representations of gaps and uncertainty for specific kinds of knowledge, and a goal management and planning system for setting and achieving learning goals.**

***Index Terms*—robotics, robot learning, architectures, representations**

I. Introduction

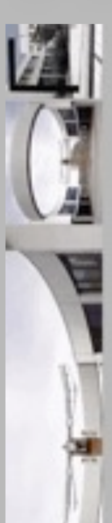
WHAT is needed for an agent to learn in a truly autonomous fashion? Autonomous learning requires that the agent pick its own learning goals. One way to achieve this is to give that agent representations of what it knows and doesn't know, and to make it reason with these representations to set its own *epistemic goals*. An epistemic goal is a goal to be in a certain knowledge state. This paper describes this approach to autonomous mental development. We present an architecture, together with a set of representations that explicitly capture what the robot and other agents do and don't know at any time, i.e. representations of their epistemic state. We also describe representations of how this epistemic state will change under action. Such representations, together with algorithms for reasoning about them confer a degree of *self-understanding*, and allow the agent to plan how to extend its abilities, or knowledge of the environment, i.e. *self-extension*. We also describe a goal management system that allows the robot to choose quickly between different epistemic goals. This mechanism is necessary to allow our approach to scale, since if a robot generates many possible learning goals the time taken to plan for them all will be too great.

Jeremy L. Wyatt, Marc Hanheide and Nick Hawes are with the University of Birmingham, email: {jlw,nah,m.hanheide}@cs.bham.ac.uk
Michael Brenner is with Albert-Ludwigs-Universität, email: brenner@informatik.uni-freiburg.de
Pierre Lison, Geert-Jan M. Kruijff and Hendrik Zender are with DFKI, Saarbrücken, Germany, email: {plison,gj,zender}@dfki.de
Patric Jensfelt, Andrzej Pronobis, Kristoffer Sjöö and Alper Aydemir are with KTH Stockholm, email: {patric,krsj,pronobis,aydemir}@csc.kth.se
Matej Kristan, Alen Vrečko and Danijel Skočaj are with University of Ljubljana, email: {matej.kristan,alen.vrecko,danijel.skocaj}@fri.uni-lj.si
Michael Zillich is with Vienna University of Technology, email: zillich@acin.tuwien.ac.at
Manuscript received February 28, 2010, revised Sept 13, 2010

We first define self-understanding and self-extension as we see them. To do this it is necessary to characterise the different types of incompleteness in knowledge that will be represented. We use *incompleteness* as an umbrella term to cover many different types of *knowledge gaps* and *uncertainty about knowledge*. We can construct a typology of incompleteness in knowledge based on three dimensions of variability. These are *the nature of the incompleteness*, the *type of knowledge that is incomplete*, and whether the incompleteness is represented in a *quantitative or qualitative* manner.

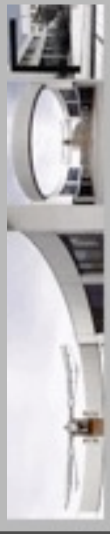
With regard to the nature of the incompleteness, in the simplest case we may have a variable or variables that are part of a model of the world and which have a defined set of possible values or hypotheses from which the true value is known to be drawn. We refer to this as *simple uncertainty*. We can also have *uncertainty about the number of variables* needed in a model, i.e. about the model complexity. Finally we can also have cases where the agent knows that a variable is of an unexperienced class, i.e. there is *novelty*. This can include cases where the variables are continuous but where the observation models for a class are quite confident and do not generalise well to some new observation. The type of knowledge that is incomplete may vary enormously. Five simple types that cover a variety of cases include contingent knowledge about the current world *state*, *structural* knowledge about the universal relationships between variables, *procedural* knowledge about how to act in certain situations to achieve certain goals, knowledge consisting of *predictions* of action outcomes or events, and knowledge about their *causes*. Finally there is a question about whether the representation is qualitative or quantitative. In qualitative representations we simply have a set of possible values for the variable, or a statement that the variable value is unknown, or knowledge that there may be many variables that are unmodelled. In quantitative representations we will have some kind of scalar values attached to hypotheses (such as whether there is novelty or not), and in our case these will typically be probabilities. Note that by a *quantitative gap* or *quantitative uncertainty* we do not mean that the underlying space for the variable is continuous or discrete, but instead that the way the incompleteness is represented involves an expression of a degree of preference for one hypothesis or statement versus another.

Given this brief characterisation of different types of incompleteness we can define self-understanding and self-extension compactly as follows. A system with *self-understanding* is any system with explicit representations that captures some



Example: Conceptual mapping

- Problem setting
  - How to act in, and talk about, dynamic large-scale environments ~ about which a robot only ever has partial knowledge?
- Development
  - Use of human-like ontology of common sense indoor knowledge,
  - connected to low-level mapping,
  - and dialogue meaning,
  - to talk, and reason, about structure and content of indoor spaces



Example: Conceptual mapping

- Problem setting
  - How to act in, and talk about, dynamic large-scale environments ~ about which a robot only ever has partial knowledge?
- Development
  - Use of human-like ontology of common sense indoor knowledge,
  - connected to low-level mapping,
  - and dialogue meaning,
  - to talk, and reason, about structure and content of indoor spaces

Situated Dialogue and Spatial Organization: What, Where… and Why?

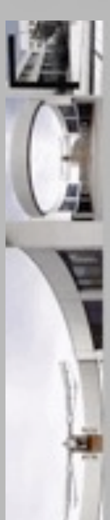
Geert-Jan M. Kruijff[1]; Hendrik Zender[1]; Patric Jensfelt[2] & Henrik I. Christensen[2]
[1]Language Technology Lab, German Research Center for Artificial Intelligence (DFKI GmbH), Saarbrücken, Germany
[2]Centre for Autonomous Systems, Royal Institute of Technology (KTH), Stockholm, Sweden
gj@dfki.de

Abstract: *The paper presents an HRI architecture for human-augmented mapping, which has been implemented and tested on an autonomous mobile robotic platform. Through interaction with a human, the robot can augment its autonomously acquired metric map with qualitative information about locations and objects in the environment. The system implements various interaction strategies observed in independently performed Wizard-of-Oz studies. The paper discusses an ontology-based approach to multi-layered conceptual spatial mapping that provides a common ground for human-robot dialogue. This is achieved by combining acquired knowledge with innate conceptual commonsense knowledge in order to infer new knowledge. The architecture bridges the gap between the rich semantic representations of the meaning expressed by verbal utterances on the one hand and the robot's internal sensor-based world representation on the other. It is thus possible to establish reference to spatial areas in a situated dialogue between a human and a robot about their environment. The resulting conceptual descriptions represent qualitative knowledge about locations in the environment that can serve as a basis for achieving a notion of situational awareness.*

Keywords: *Human-Robot Interaction, Conceptual Spatial Mapping, Situated Dialogue*

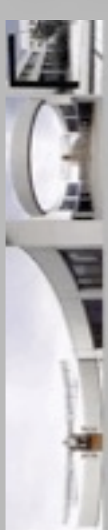
1. Introduction

More and more robots find their way into environments where their primary purpose is to interact with humans to help and solve a variety of service-oriented tasks. Particularly if such a service robot is mobile, it needs to have an understanding of the spatial and functional properties of the environment in which it operates. The problem we address is how a robot can acquire an understanding of the environment so that it can autonomously operate in it, and communicate about it with a human. We present an architecture that provides the robot with this ability through a combination of human-robot interaction and autonomous mapping techniques. It captures various functions that independently performed Wizard-of-Oz studies have observed to be necessary for such a system. Several case studies have been conducted to test and evaluate the resulting integrated system.
The main issue is how to establish a correspondence between how a human perceives spatial and functional aspects of an environment, and what the robot autonomously learns as a map. Most existing approaches to robot map building, or Simultaneous Localization And Mapping (SLAM), use a metric representation of space. Humans, though, have a more qualitative, topological perspective on spatial organization (McNamara, 1986). We adopt an approach in which we build a multi-layered representation of the environment, combining metric maps and topological graphs (as an abstraction over geometrical information), like (Kuipers, 2000). We extend these representations with conceptual descriptions that capture aspects of spatial and functional organization. The robot obtains these descriptions either through interaction with a human, or through inference combining its own observations (*I see a coffee machine*) with ontological knowledge (*Coffee machines are usually found in kitchens, so this is likely to be a kitchen!*). We store objects in the spatial representations, and so associate the functionality of a location with that of the functions of the objects present there. A core characteristic of our approach is that we analyze each utterance to obtain a representation of the meaning it expresses, and how it (syntactically) conveys that meaning – rather than just doing for example keyword spotting. This way, we can properly handle the variety of ways in which people may express assertions, questions, and commands. Furthermore, having a representation of the meaning of the utterance we can combine it with further inferences over ontologies to obtain a complete conceptual description of the location or object being talked about. This way we can ground situated dialogue in the situational awareness of the robot.
Following (Topp & Christensen, 2005) and (Topp et al., 2006), we talk about *Human-Augmented Mapping* (HAM) to indicate the active role that human-robot interaction plays in the robot's acquisition of qualitative spatial knowledge. In §2 we discuss various observations that independently performed Wizard-of-Oz studies have made on typical interactions for HAM scenarios, and we



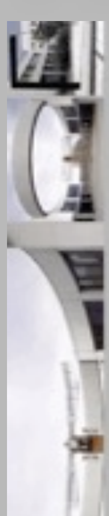
Example: Conceptual mapping (2006)

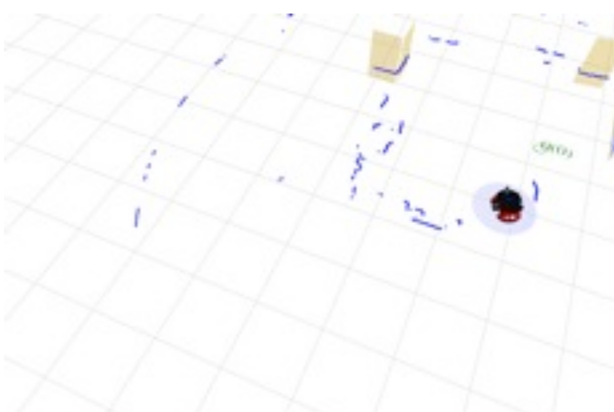
Structured environments in which meaning is
"given," "fixed" by human interpretation



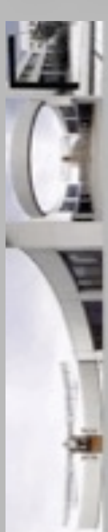
Example: Conceptual mapping (2006)

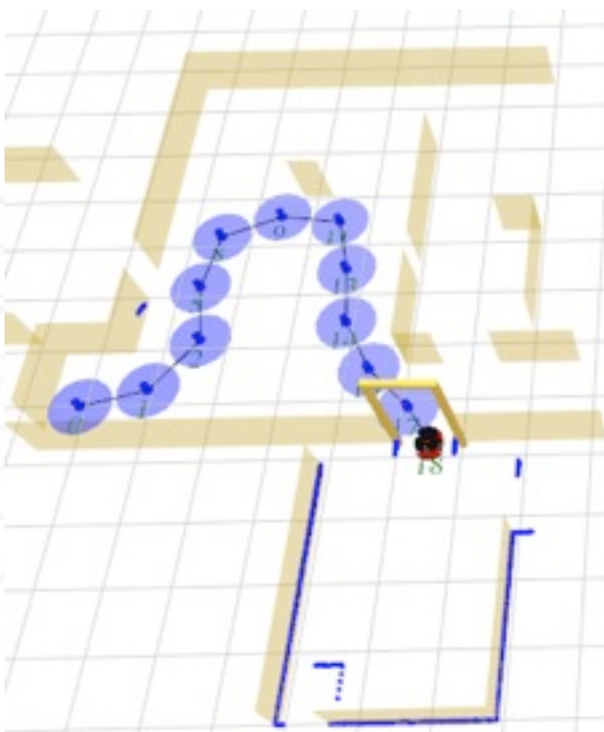
Structured environments in which meaning is “given,” “fixed” by human interpretation



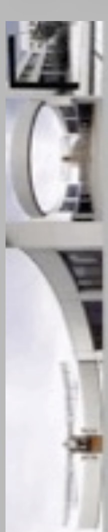


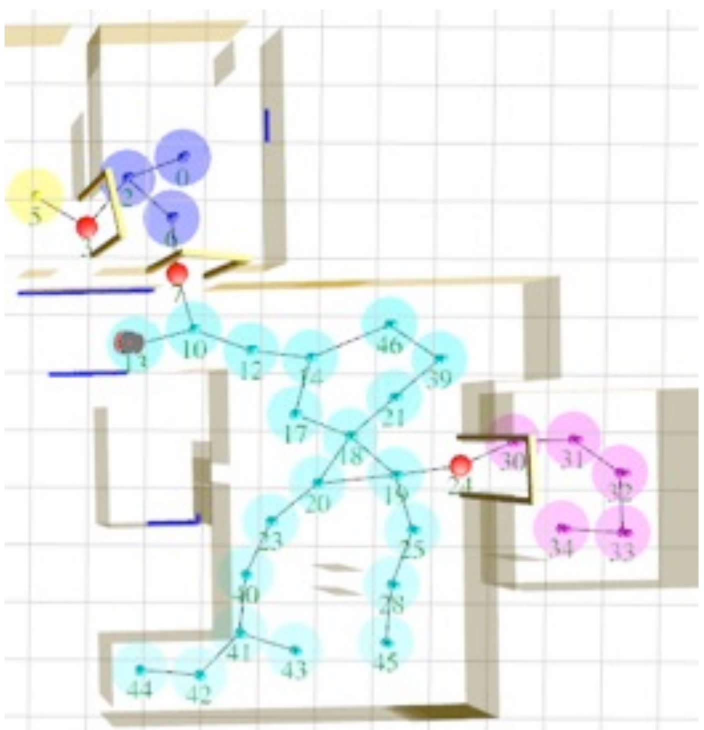
Layered (cross-)modal beliefs



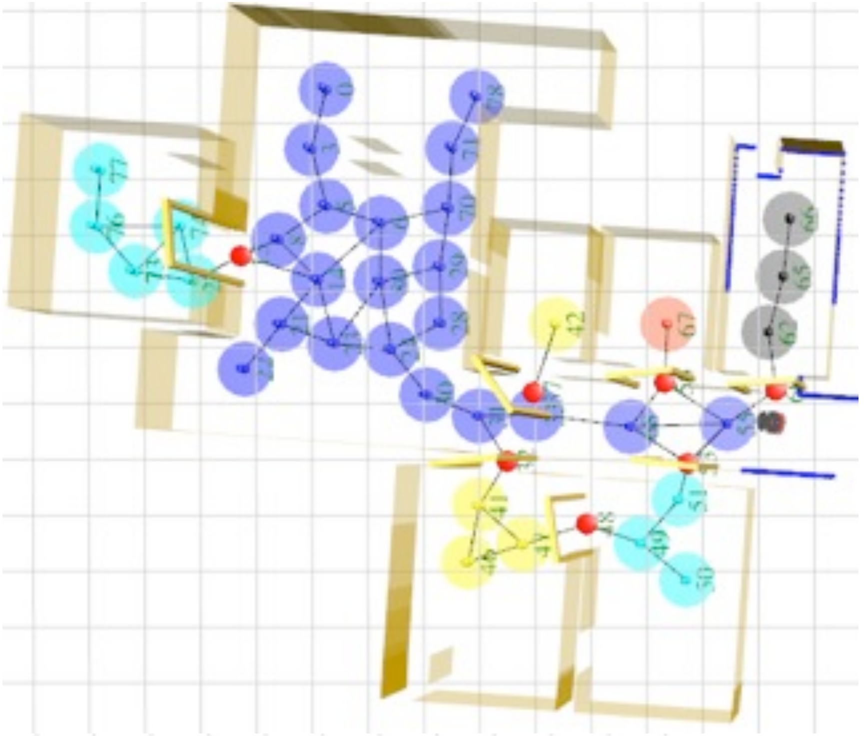


Layered (cross-)modal beliefs

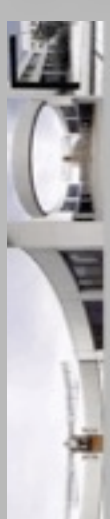


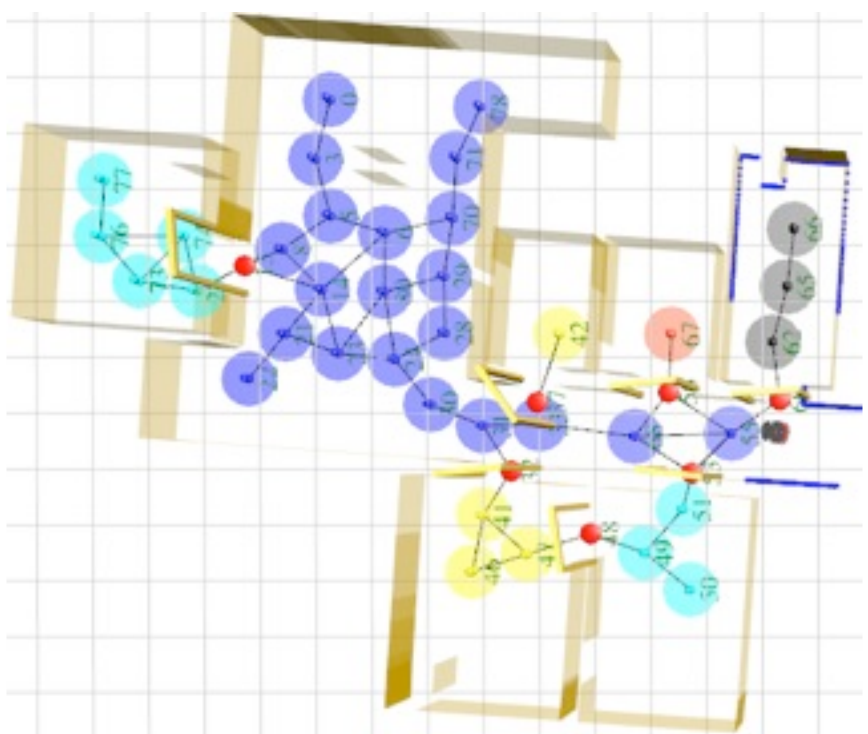


Layered (cross-)modal beliefs

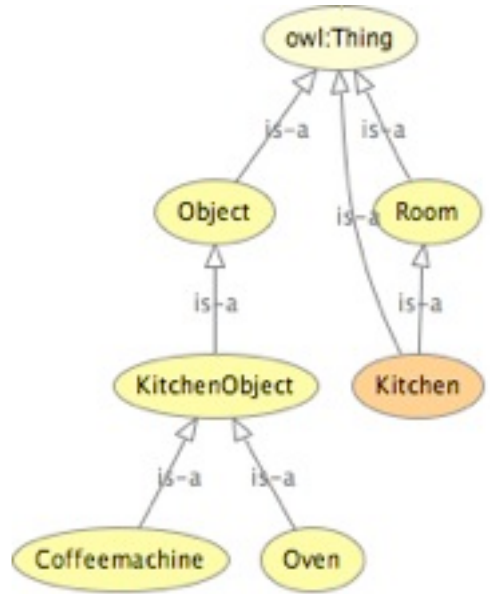


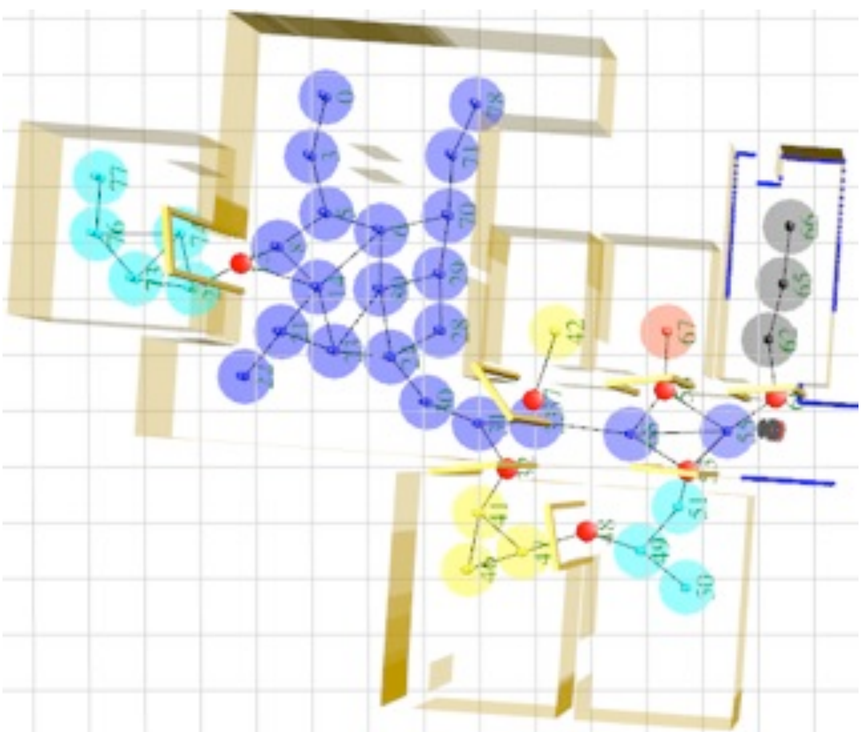
Layered (cross-)modal beliefs



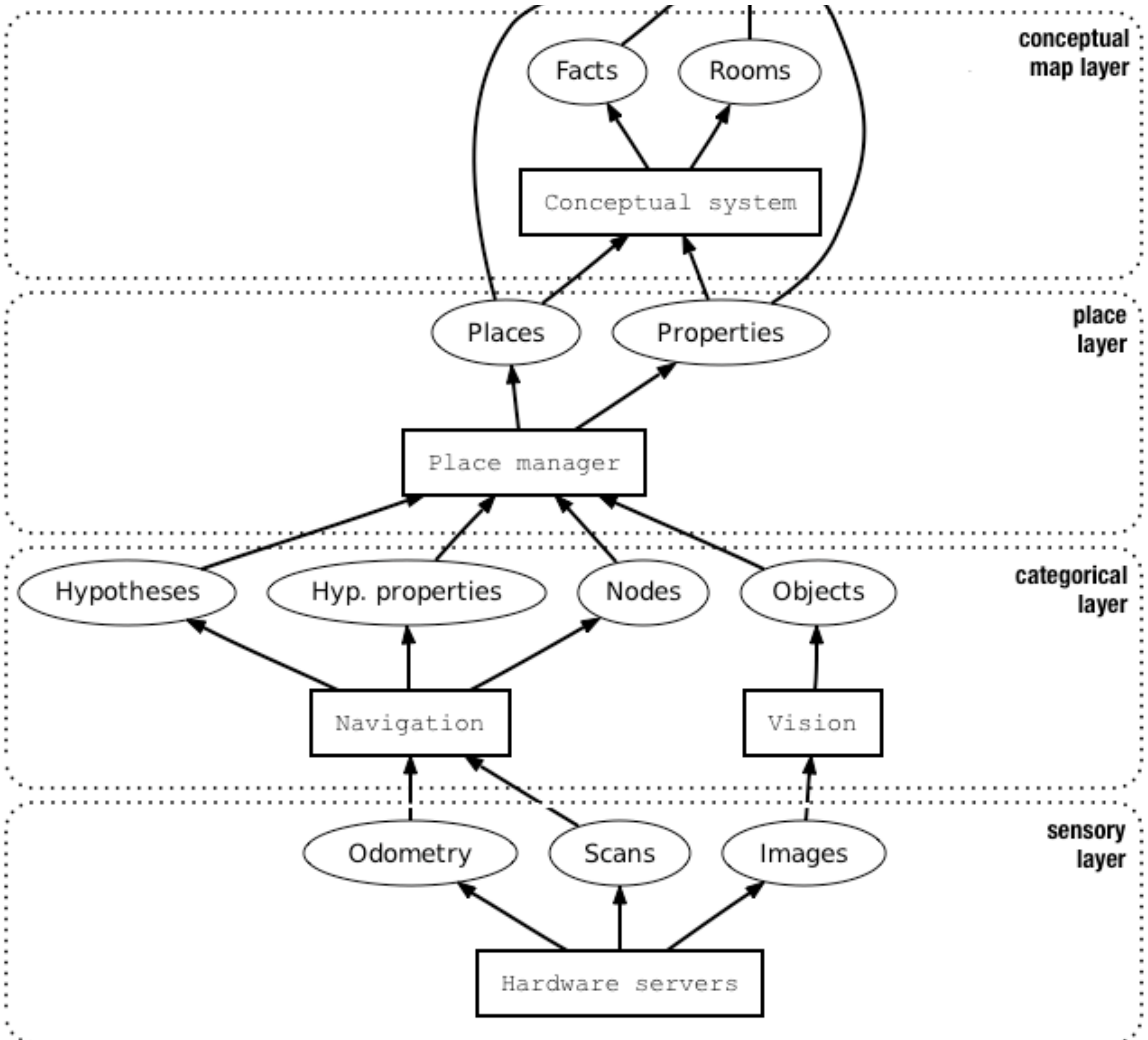


Layered (cross-)modal beliefs





Layered (cross-)modal beliefs

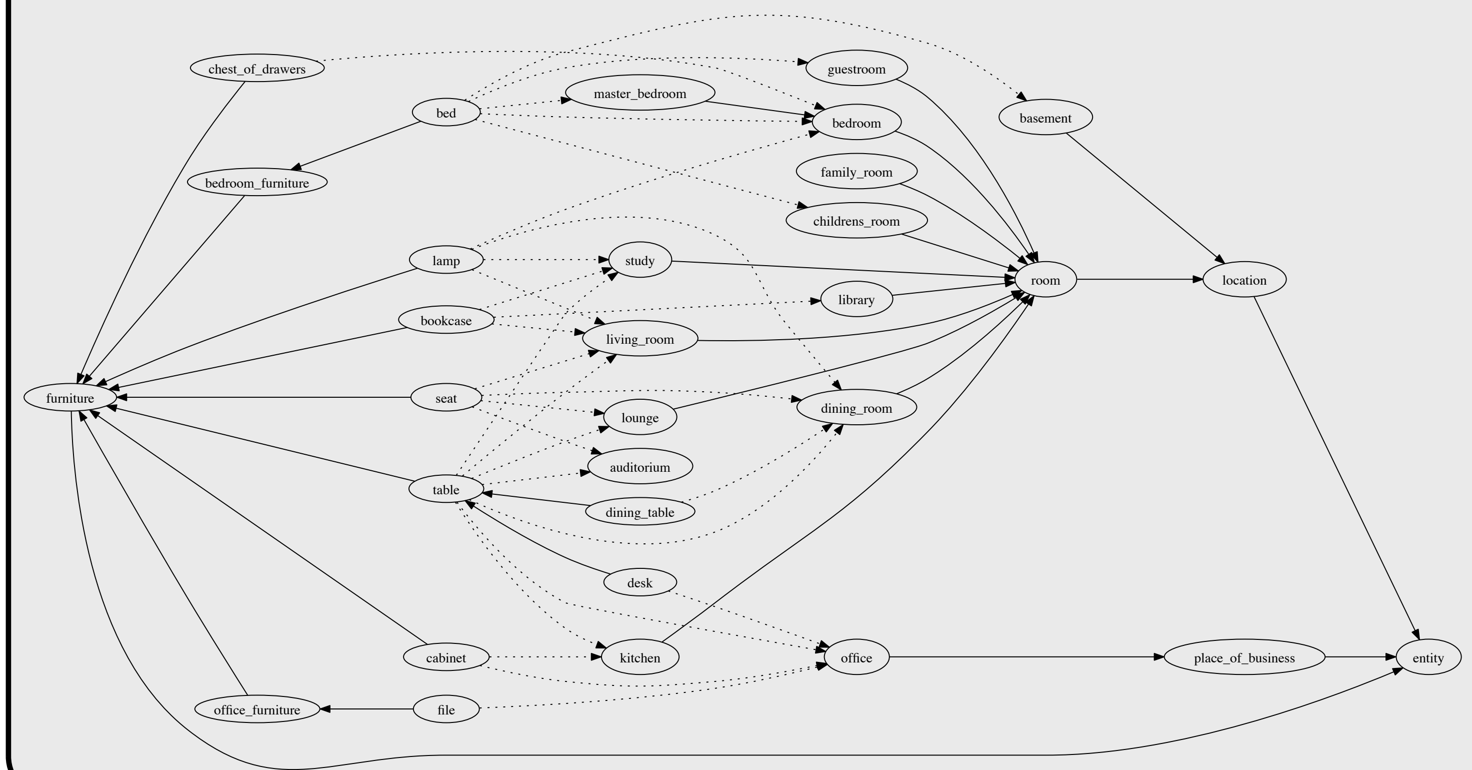


Layered (cross-)modal beliefs

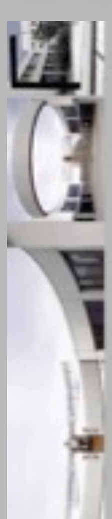
OPEN MIND
indoor common sense

PRINCETON UNIVERSITY
WordNet
A lexical database for English

	kitchen	bathroom	garage	office
sink	0.394958	0.24747899	0.053361345	0.05630252
faucet	0.45874125	0.40419582	0.018181818	0.043776225
computer	0.048387095	0.028830646	0.019112904	0.111693546

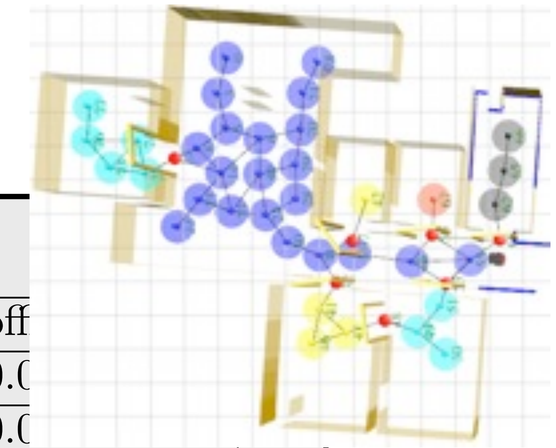


Referencing

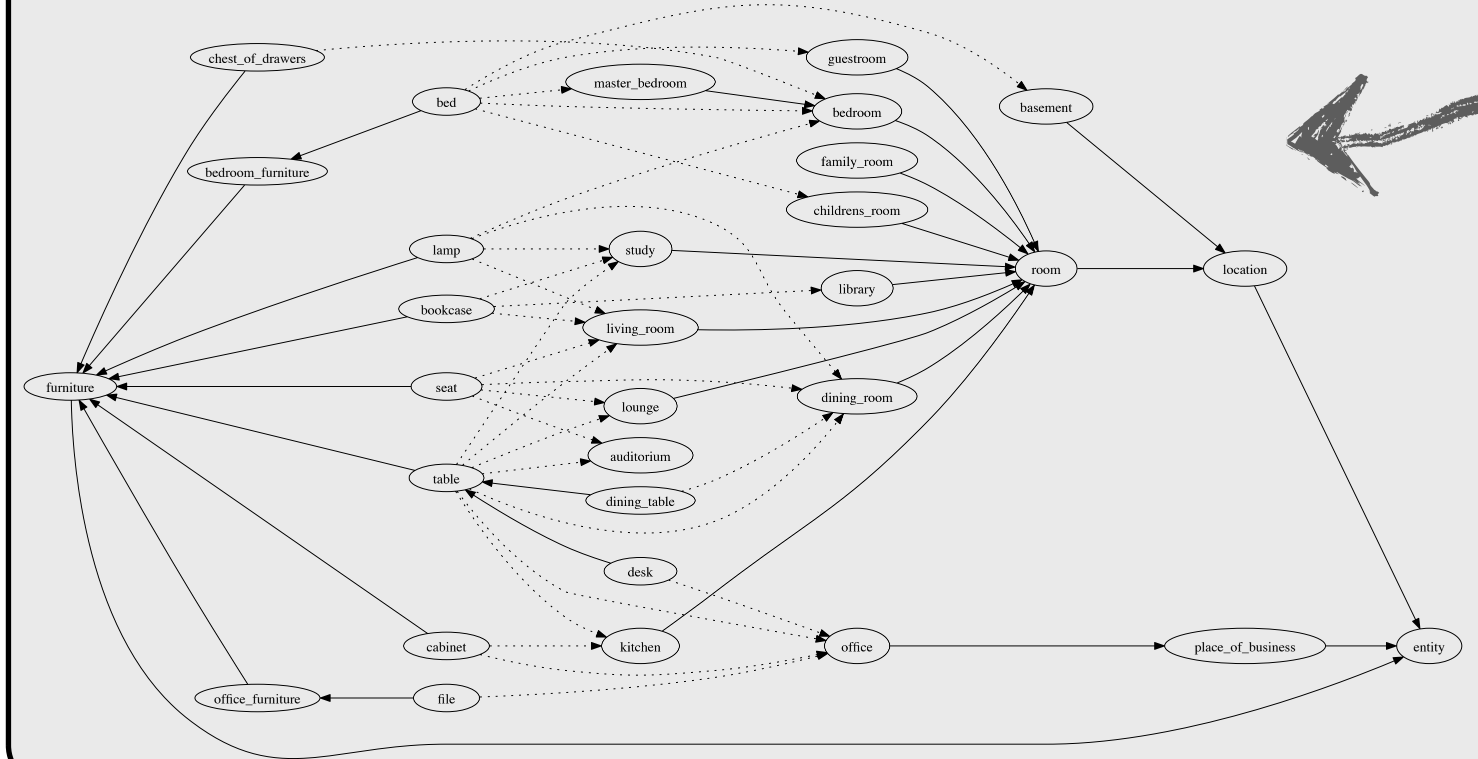


OPEN MIND
indoor common sense

PRINCETON UNIVERSITY
WordNet
A lexical database for English

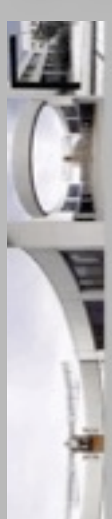


	kitchen	bathroom	garage	off
sink	0.394958	0.24747899	0.053361345	0.0
faucet	0.45874125	0.40419582	0.018181818	0.0
computer	0.048387095	0.028830646	0.019112904	0.111693546



Referencing

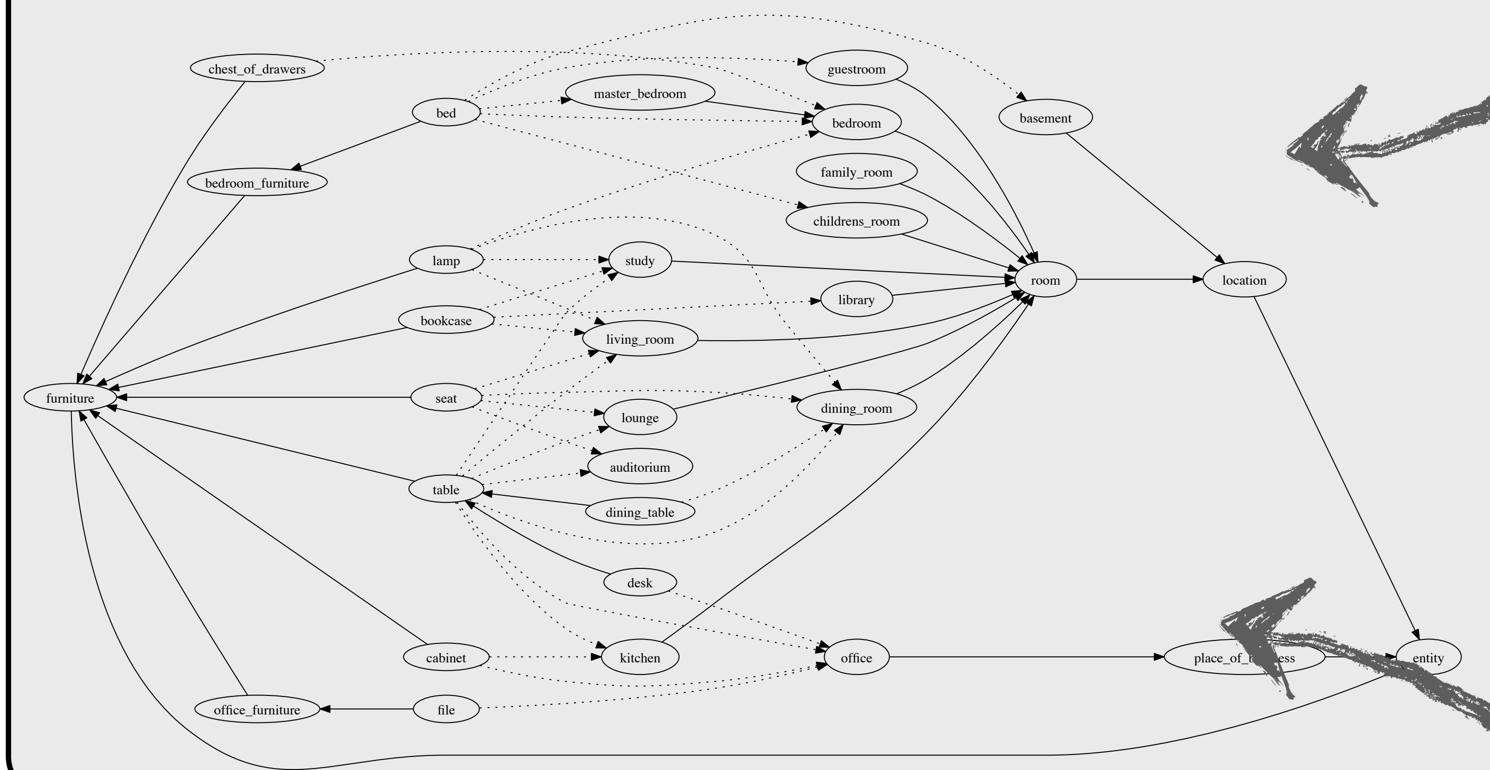
Talking Robots
Language Technology, DFKI

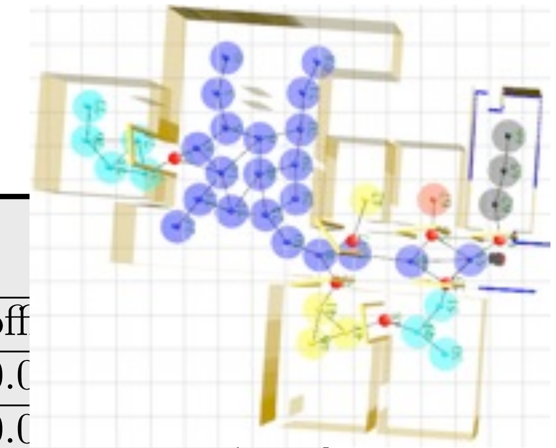


OPEN MIND
indoor common sense

PRINCETON UNIVERSITY
WordNet
A lexical database for English

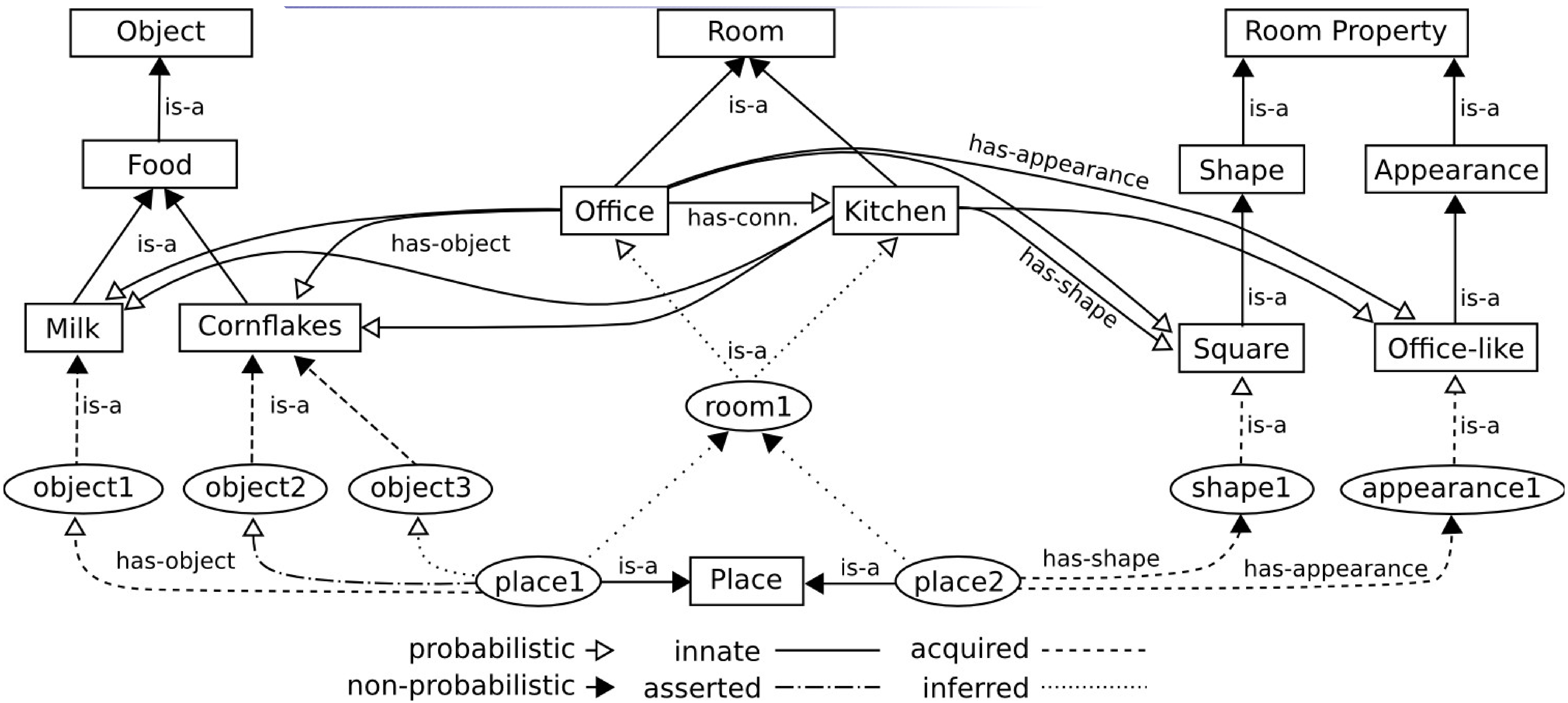
	kitchen	bathroom	garage	off
sink	0.394958	0.24747899	0.053361345	0.0
faucet	0.45874125	0.40419582	0.018181818	0.0
computer	0.048387095	0.028830646	0.019112904	0.111693546



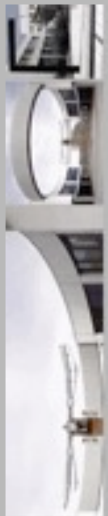


Dialogue Understanding

Referencing



Linked (ontological) beliefs with tendencies, instances, hypotheses



Example: Spatial reference

- Problem setting
  - Using conceptual map structure to resolve and produce contextually-appropriate referring expressions
- Development
  - Use of human-like ontology of common sense indoor knowledge,
  - and use of hierarchical structure in map,
  - to determine appropriate spatial context for establishing reference against in-context distractors

Example: Spatial reference

- Problem setting
    - Using conceptual map structure to resolve and produce contextually-appropriate referring expressions
- Development
    - Use of human-like ontology of common sense indoor knowledge,
    - and use of hierarchical structure in map,
    - to determine appropriate spatial context for establishing reference against in-context distractors

Situated Resolution and Generation of Spatial Referring Expressions for Robotic Assistants*

Hendrik Zender and **Geert-Jan M. Kruijff** and **Ivana Kruijff-Korbayová**
Language Technology Lab, German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
{zender, gj, ivana.kruijff}@dfki.de

Abstract

In this paper we present an approach to the task of generating and resolving referring expressions (REs) for conversational mobile robots. It is based on a spatial knowledge base encompassing both robot- and human-centric representations. Existing algorithms for the generation of referring expressions (GRE) try to find a description that uniquely identifies the referent with respect to other entities that are in the current context. Mobile robots, however, act in large-scale space, that is, environments that are larger than what can be perceived at a glance, e.g., an office building with different floors, each containing several rooms and objects. One challenge when referring to elsewhere is thus to include enough information so that the interlocutors can extend their context appropriately. We address this challenge with a method for context construction that can be used for both generating and resolving REs – two previously disjoint aspects. Our approach is embedded in a bi-directional framework for natural language processing for robots.

1 Introduction

The past years have seen an extraordinary increase in research on robotic assistants that help the users perform their daily chores. Although the autonomous vacuum cleaner "Roomba" has already found its way into people's homes and lives, there is still a long way until fully conversational robot "gophers" will be able to assist people in more demanding everyday tasks. For example, imagine a robot that can deliver objects and give directions to visitors on a university campus. Such a robot must be able to verbalize its knowledge in a way that is understandable by humans, as illustrated in Figure 1.

A conversational robot will inevitably face situations in which it needs to refer to an entity (e.g., an object, a locality, or even an event) that is located somewhere outside the current scene. There are conceivably many ways in which a robot might refer to things in the world, but many such expressions are unsuitable in most human-robot dialogues. Consider the following set of examples:

*Supported by the EU FP7 Project "CogX" (FP7-ICT-215181).

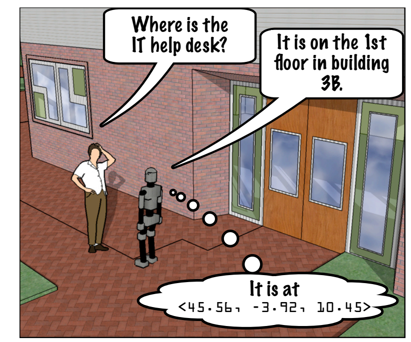


Figure 1: Situated dialogue with a campus service robot

1. "position $P = \langle 45.56, -3.92, 10.45 \rangle$"
2. "the area"
3. "Peter's office at the end of the corridor on the third floor of the Acme Corp. building 7 in the Acme Corp. complex, 47 Evergreen Terrace, Calisota, Earth, (...)"

Clearly, these REs are valid descriptions of the respective entities in the robot's world representation. Still they fail to achieve their *communicative goal*, which is to specify the right amount of information so that the hearer can easily uniquely identify what is meant. The following expressions *might* serve as more appropriate variants of the previous examples (*in certain situations!*):

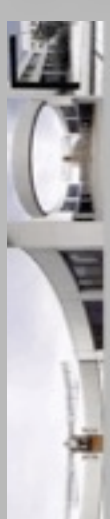
1. "the IT help desk"
2. "the large hall on the first floor"
3. "Peter's office"

However, the question remains how a natural language processing (NLP) system can generate such expressions which are suitable in a given situation. In this paper we identify some of the challenges that an NLP system for situated dialogue about large-scale space needs to address. We present a situated model for generating and resolving REs that addresses these issues, with a special focus on how a conversational mobile robot can produce and interpret such expressions against an appropriate part of its acquired knowledge base (KB). One benefit of our approach is that most components, including the situated model and the linguistic resources, are bi-directional, i.e., they use the same representa-

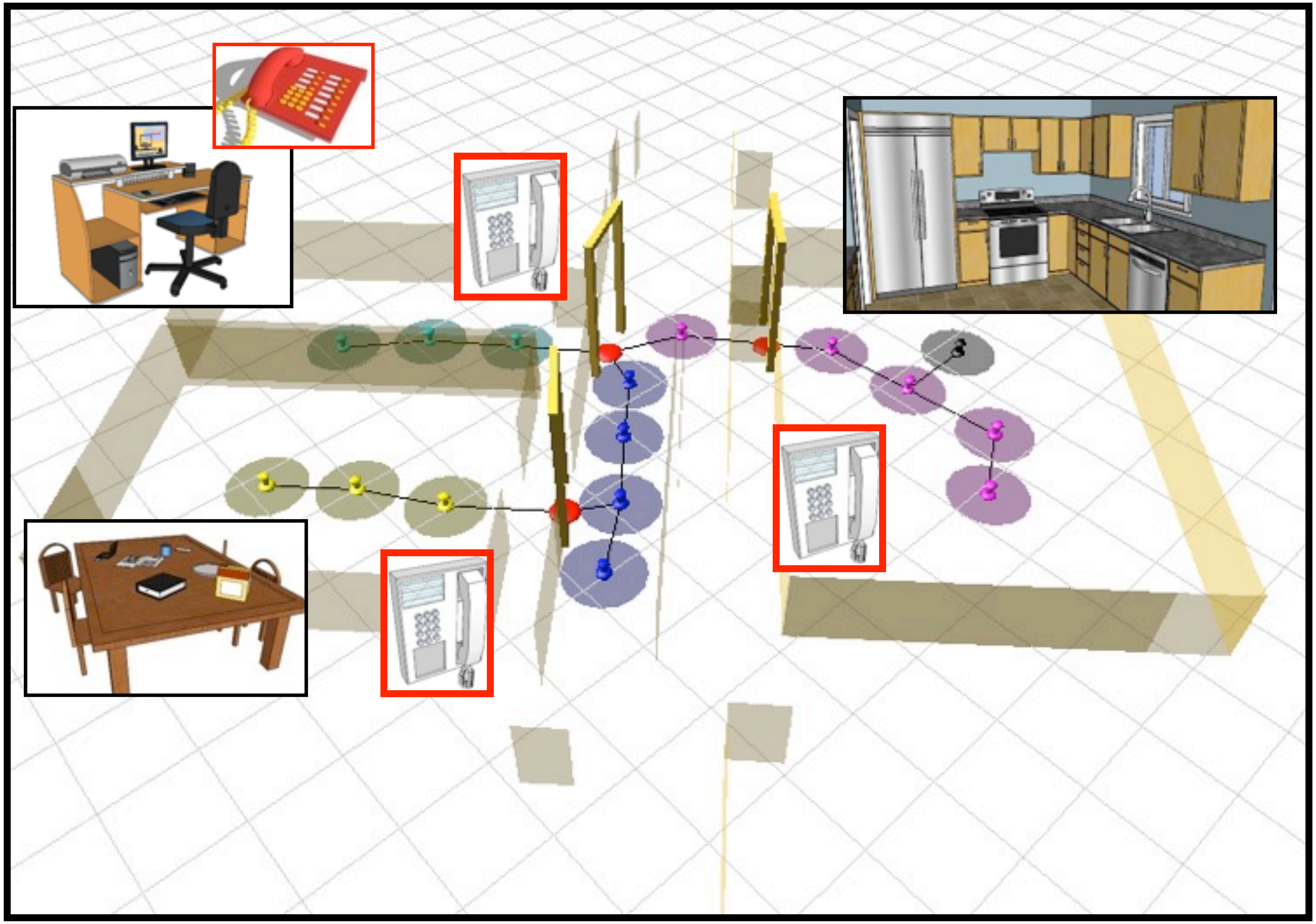
Generating referring expressions

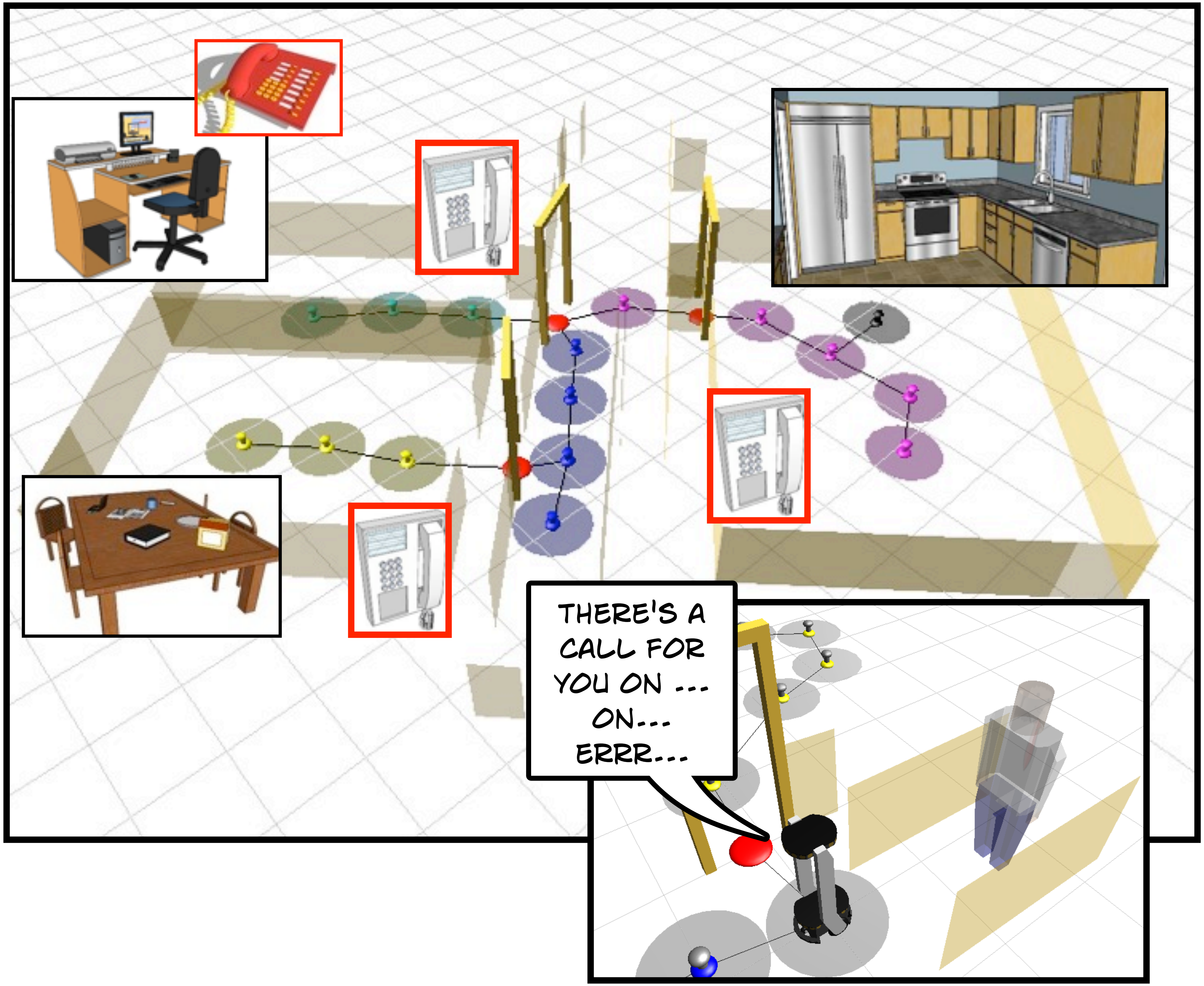
Generating referring expressions

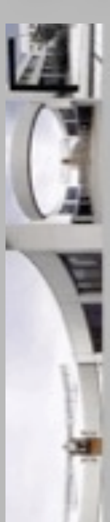
Generating referring expressions

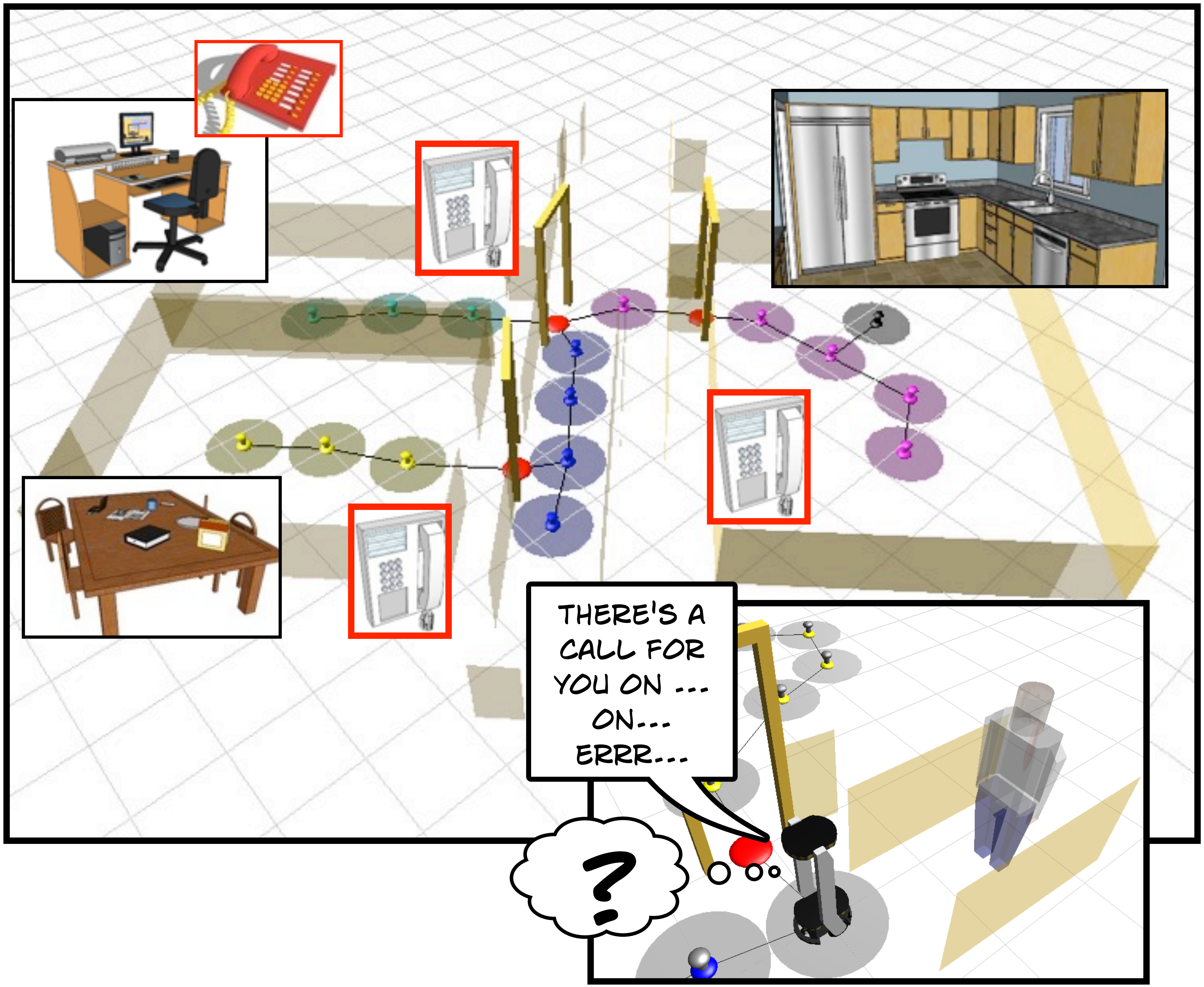


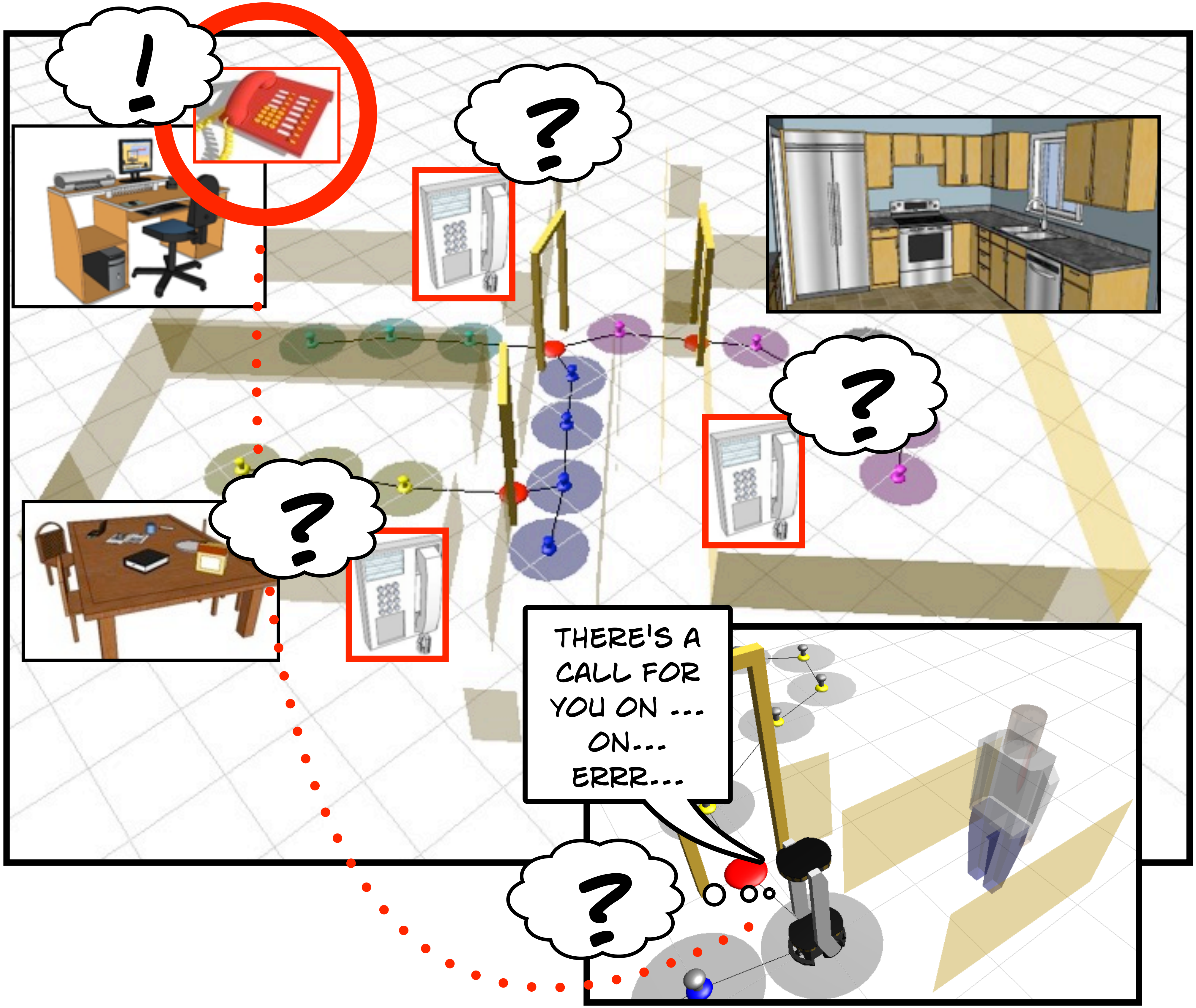
Generating referring expressions

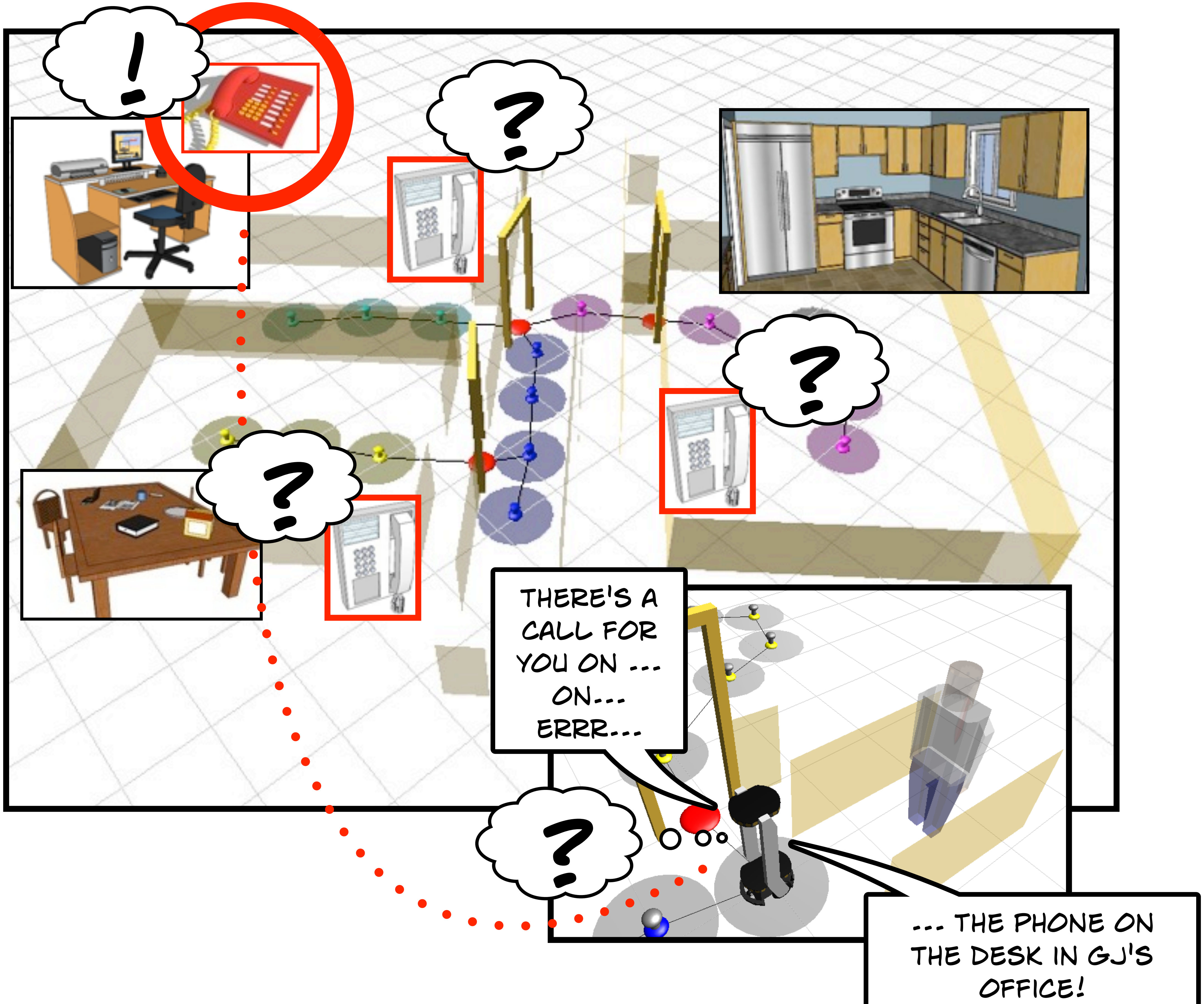



THERE'S A CALL FOR YOU ON ... ON... ERRR...



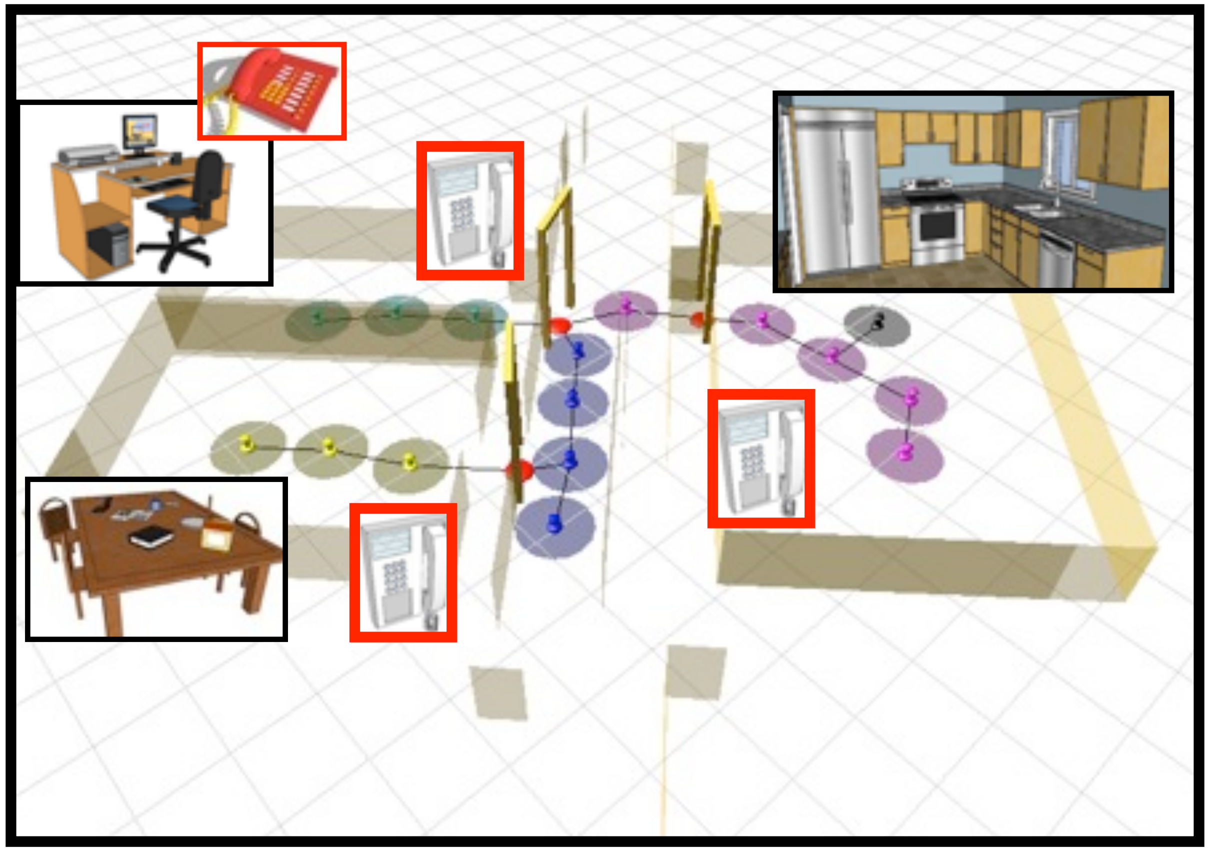

THERE'S A CALL FOR YOU ON ... ON... ERRR...
?


!
?
?
?
?
There's a call for you on ...
on...
errr...

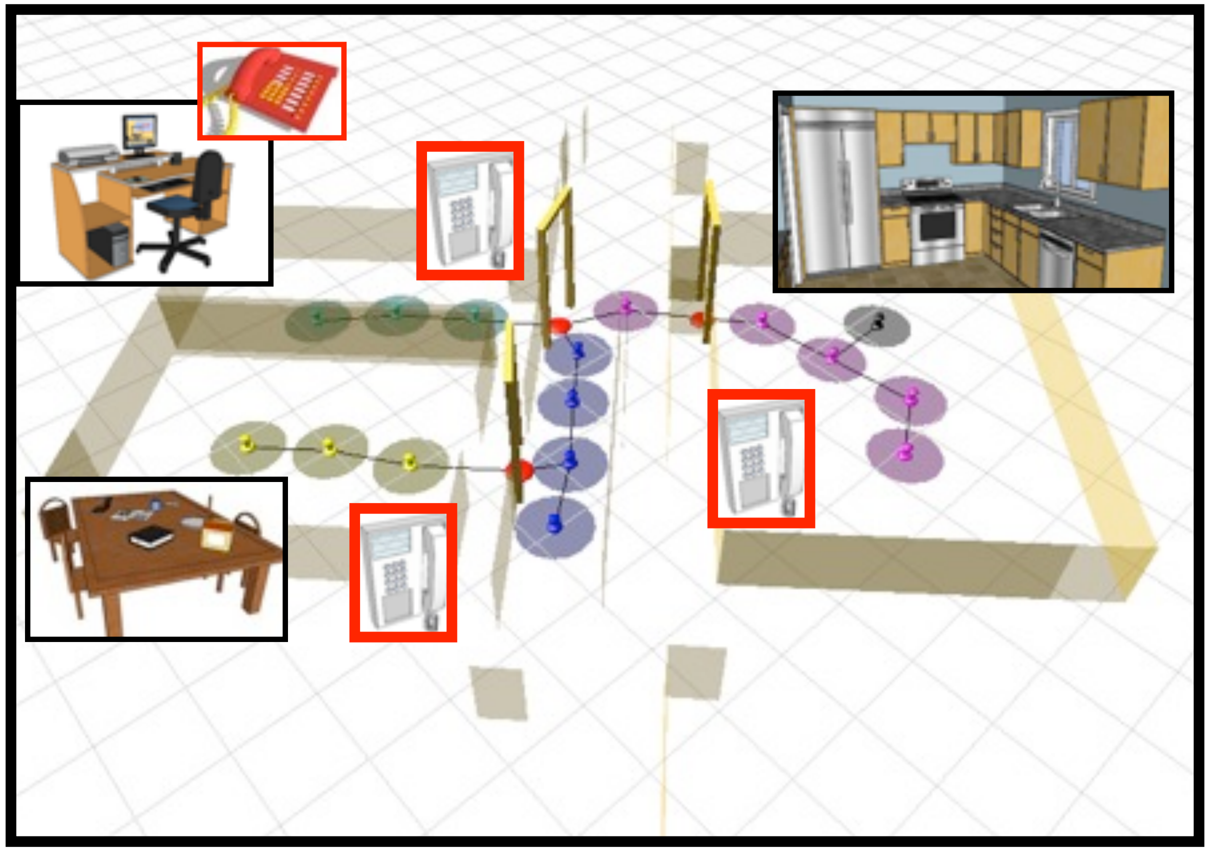


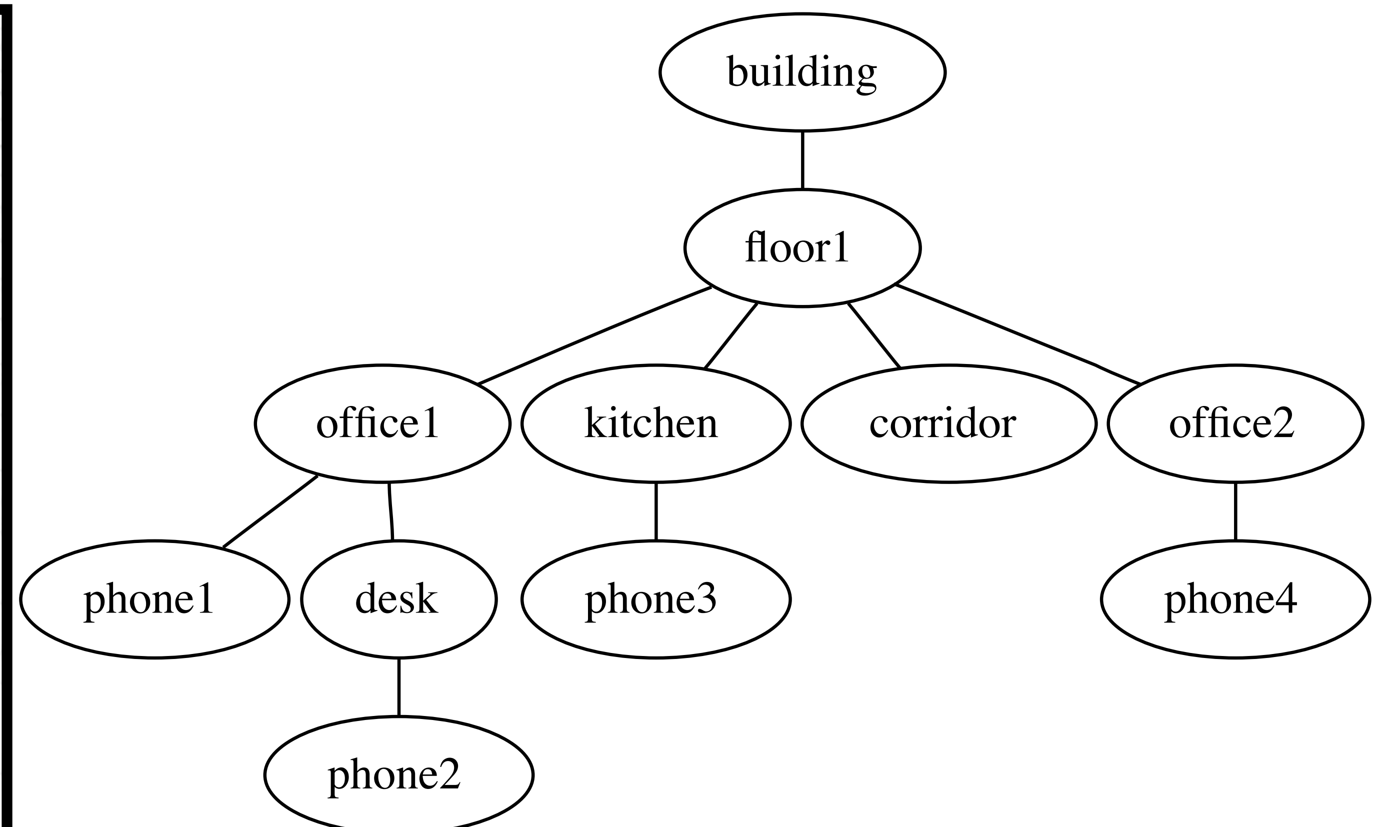
TALKING ROBOTS
LANGUAGE TECHNOLOGY, DFKI

Containment and abstraction



Containment and abstraction

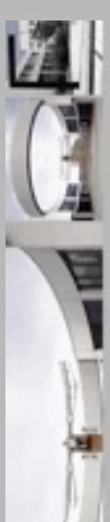




Containment and abstraction

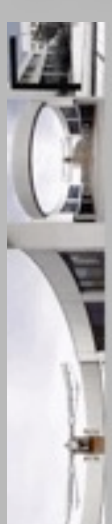
Mediation

- Ontology-based mediation
- Objects, features, uncertainty
- Reasoning with incompleteness
- Structural uncertainty



Mediation

- Ontology-based mediation
- Objects, features, uncertainty
- Reasoning with incompleteness
- Structural uncertainty



IEEE RO-MAN 2010
19th International Symposium on Robot and Human Interactive Communication
September 12-15, 2010 | Principe di Piemonte, Viareggio, Italy

Belief Modelling for Situation Awareness in Human-Robot Interaction

Pierre Lison, Carsten Ehrler and Geert-Jan M. Kruijff

***Abstract*— To interact naturally with humans, robots needs to be aware of their own surroundings. This awareness is usually encoded in some implicit or explicit representation of the situated context. In this paper, we present a new framework for constructing rich belief models of the robot's environment. Key to our approach is the use of *Markov Logic* as a unified framework for inference over these beliefs. Markov Logic is a combination of first-order logic and probabilistic graphical models. Its expressive power allows us to capture both the rich relational structure of the environment and the uncertainty arising from the noise and incompleteness of low-level sensory data. The constructed belief models evolve dynamically over time and incorporate various contextual information such as spatio-temporal framing, multi-agent epistemic status, and saliency measures. Beliefs can also be referenced and extended "top-down" via linguistic communication. The approach is being integrated into a cognitive architecture for mobile robots interacting with humans using spoken dialogue.**

I. Introduction

The situated context plays a central role in human-robot interaction (HRI). To be able to interact naturally with humans, robots needs to be aware of their own environment. This situation awareness is generally expressed in some sort of *belief models* in which various aspects of the external reality are encoded. Such belief models provide an explicit or implicit representation for the current state of the world, from the robot's viewpoint. They therefore serve as a representational backbone for a wide range of high-level cognitive capabilities related to reasoning, planning and learning in complex and dynamic environments. They are also essential for the robot to verbalise its own knowledge.

In speech-based HRI, critical tasks in dialogue understanding, management and production are directly dependent on such belief models. Examples are context-sensitive speech recognition [15], reference resolution and generation in small- [11] and large-scale space [24], spoken dialogue parsing [14] and interpretation [20], dialogue management [23], user-tailored response generation [22], and contextually appropriate intonation patterns [13]. Contextual knowledge is also a prerequisite for the dynamic adaptation of the robot's behaviour to different environments and interlocutors [3].

Belief models are usually expressed as high level symbolic representations merging and abstracting information over multiple modalities. For HRI, the incorporated knowledge might include (inter alia): entities in the visual scene, spatial structure, user profiles (intentional and attentional state, preferences), dialogue histories, and task models (what is to be done, which actions are available).

The construction of such belief models raises two important issues. The first question to address is how these high-level representations can be reliably abstracted from low-level sensory data [1], [18]. To be meaningful, most symbolic representations should be *grounded* in (subsymbolic) sensory inputs [19]. This is a difficult problem, partly because of the noise and uncertainty contained in sensory data (partial observability), and partly because the connection between low-level perception and high-level symbols is typically difficult to formalise in a general way [6].

The second issue relates to how information arising from different modalities and time points can be efficiently *merged* into unified multi-modal structures [12], and how these inputs can refine and constrain each other to yield improved estimations, over time. This is the well-known engineering problem of multi-target, multi-sensor data fusion [5].

Belief models are thus the final product of an iterative process of information *fusion*, *refinement* and *abstraction*. Typical HRI environments are challenging to model, being simultaneously *complex*, *multi-agent*, *dynamic* and *uncertain*. Four requirements can be formulated:

1) HRI environments are complex and reveal a large amount of internal structure (for instance, spatial relations between entities, or groupings of objects). The formal representations used to model them must therefore possess the expressive power to reflect this rich relational structure.
2) Interactive robots are made for multi-agent settings. Making sense of communicative acts requires the ability to distinguish between one's own knowledge (what I believe), knowledge attributed to others (what I think the others believe), and shared common ground knowledge (what we believe as a group).
3) Situated interactions are *dynamic* and evolve over time. The incorporation of spatio-temporal framing is thus necessary to go beyond the "here-and-now" and be capable of linking the present with (episodic) memories of the past and anticipation of future events.
4) And last but not least, due to the partial observability of most contextual features, it is crucial that belief models incorporate an explicit account of *uncertainties*.

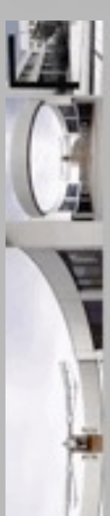
Orthogonal to these "representational" requirements, crucial performance requirements must also be adressed. To keep up with a continuously changing environment, all operations on belief models (updates, queries, etc.) must be performed under soft real-time constraints.

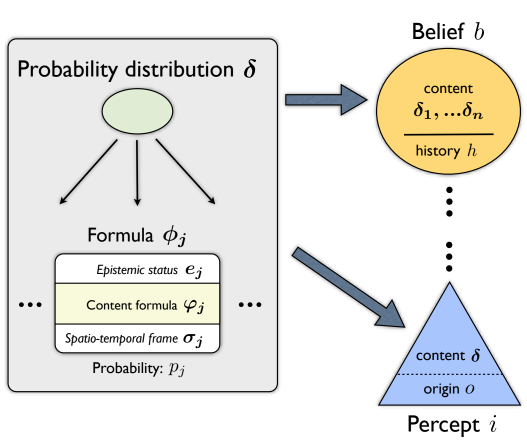
This work was supported by the EU FP7 ICT Integrated Project "*CogX: cognitive systems that self-understand and self-extend*" (FP7-ICT- 215181).

Pierre Lison, Carsten Ehrler and Geert-Jan M. Kruijff are with the German Research Centre for Artificial Intelligence (DFKI GmbH), Language Technology Lab, Saarbrücken, Germany. {plison,carsten.ehrler,gj}@dfki.de

Mediation

- Ontology-based mediation
- Objects, features, uncertainty
- Reasoning with incompleteness
- Structural uncertainty





IEEE RO-MAN 2010
19th International Symposium on Robot and Human Interactive Communication
September 12-15, 2010 | Principe di Piemonte, Viareggio, Italy

Belief Modelling for Situation Awareness in Human-Robot Interaction

Pierre Lison, Carsten Ehrler and Geert-Jan M. Kruijff

***Abstract*— To interact naturally with humans, robots needs to be aware of their own surroundings. This awareness is usually encoded in some implicit or explicit representation of the situated context. In this paper, we present a new framework for constructing rich belief models of the robot's environment. Key to our approach is the use of *Markov Logic* as a unified framework for inference over these beliefs. Markov Logic is a combination of first-order logic and probabilistic graphical models. Its expressive power allows us to capture both the rich relational structure of the environment and the uncertainty arising from the noise and incompleteness of low-level sensory data. The constructed belief models evolve dynamically over time and incorporate various contextual information such as spatio-temporal framing, multi-agent epistemic status, and saliency measures. Beliefs can also be referenced and extended "top-down" via linguistic communication. The approach is being integrated into a cognitive architecture for mobile robots interacting with humans using spoken dialogue.**

I. Introduction

The situated context plays a central role in human-robot interaction (HRI). To be able to interact naturally with humans, robots needs to be aware of their own environment. This situation awareness is generally expressed in some sort of *belief models* in which various aspects of the external reality are encoded. Such belief models provide an explicit or implicit representation for the current state of the world, from the robot's viewpoint. They therefore serve as a representational backbone for a wide range of high-level cognitive capabilities related to reasoning, planning and learning in complex and dynamic environments. They are also essential for the robot to verbalise its own knowledge.

In speech-based HRI, critical tasks in dialogue understanding, management and production are directly dependent on such belief models. Examples are context-sensitive speech recognition [15], reference resolution and generation in small- [11] and large-scale space [24], spoken dialogue parsing [14] and interpretation [20], dialogue management [23], user-tailored response generation [22], and contextually appropriate intonation patterns [13]. Contextual knowledge is also a prerequisite for the dynamic adaptation of the robot's behaviour to different environments and interlocutors [3].

Belief models are usually expressed as high level symbolic representations merging and abstracting information over multiple modalities. For HRI, the incorporated knowledge might include (inter alia): entities in the visual scene, spatial structure, user profiles (intentional and attentional state, preferences), dialogue histories, and task models (what is to be done, which actions are available).

The construction of such belief models raises two important issues. The first question to address is how these high-level representations can be reliably abstracted from low-level sensory data [1], [18]. To be meaningful, most symbolic representations should be *grounded* in (subsymbolic) sensory inputs [19]. This is a difficult problem, partly because of the noise and uncertainty contained in sensory data (partial observability), and partly because the connection between low-level perception and high-level symbols is typically difficult to formalise in a general way [6].

The second issue relates to how information arising from different modalities and time points can be efficiently *merged* into unified multi-modal structures [12], and how these inputs can refine and constrain each other to yield improved estimations, over time. This is the well-known engineering problem of multi-target, multi-sensor data fusion [5].

Belief models are thus the final product of an iterative process of information *fusion*, *refinement* and *abstraction*. Typical HRI environments are challenging to model, being simultaneously *complex*, *multi-agent*, *dynamic* and *uncertain*. Four requirements can be formulated:

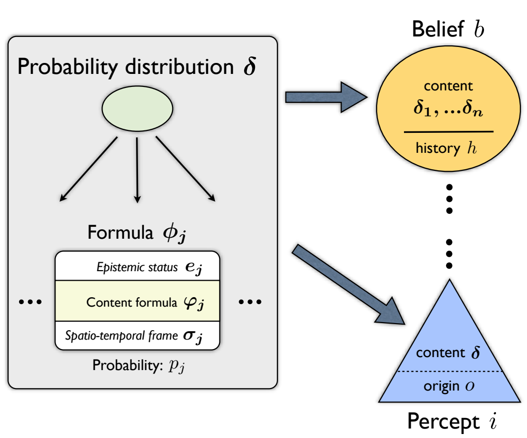
1) HRI environments are complex and reveal a large amount of internal structure (for instance, spatial relations between entities, or groupings of objects). The formal representations used to model them must therefore possess the expressive power to reflect this rich relational structure.
2) Interactive robots are made for multi-agent settings. Making sense of communicative acts requires the ability to distinguish between one's own knowledge (what I believe), knowledge attributed to others (what I think the others believe), and shared common ground knowledge (what we believe as a group).
3) Situated interactions are *dynamic* and evolve over time. The incorporation of spatio-temporal framing is thus necessary to go beyond the "here-and-now" and be capable of linking the present with (episodic) memories of the past and anticipation of future events.
4) And last but not least, due to the partial observability of most contextual features, it is crucial that belief models incorporate an explicit account of *uncertainties*.

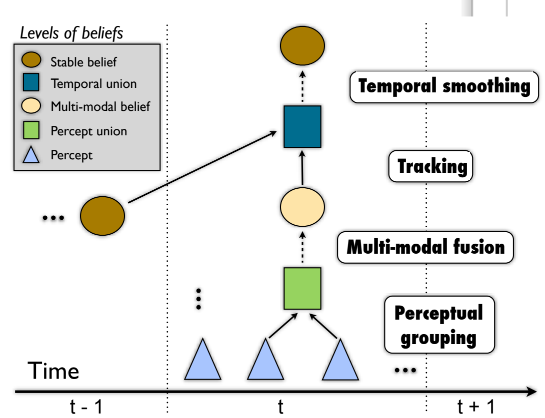
Orthogonal to these "representational" requirements, crucial performance requirements must also be adressed. To keep up with a continuously changing environment, all operations on belief models (updates, queries, etc.) must be performed under soft real-time constraints.

This work was supported by the EU FP7 ICT Integrated Project "*CogX: cognitive systems that self-understand and self-extend*" (FP7-ICT- 215181).
Pierre Lison, Carsten Ehrler and Geert-Jan M. Kruijff are with the German Research Centre for Artificial Intelligence (DFKI GmbH), Language Technology Lab, Saarbrücken, Germany. {plison,carsten.ehrler,gj}@dfki.de

Mediation

- Ontology-based mediation
- Objects, features, uncertainty
- Reasoning with incompleteness
- Structural uncertainty





Belief Modelling for Situation Awareness in Human-Robot Interaction

Pierre Lison, Carsten Ehrler and Geert-Jan M. Kruijff

***Abstract*— To interact naturally with humans, robots needs to be aware of their own surroundings. This awareness is usually encoded in some implicit or explicit representation of the situated context. In this paper, we present a new framework for constructing rich belief models of the robot's environment. Key to our approach is the use of *Markov Logic* as a unified framework for inference over these beliefs. Markov Logic is a combination of first-order logic and probabilistic graphical models. Its expressive power allows us to capture both the rich relational structure of the environment and the uncertainty arising from the noise and incompleteness of low-level sensory data. The constructed belief models evolve dynamically over time and incorporate various contextual information such as spatio-temporal framing, multi-agent epistemic status, and saliency measures. Beliefs can also be referenced and extended "top-down" via linguistic communication. The approach is being integrated into a cognitive architecture for mobile robots interacting with humans using spoken dialogue.**

I. Introduction

The situated context plays a central role in human-robot interaction (HRI). To be able to interact naturally with humans, robots needs to be aware of their own environment. This situation awareness is generally expressed in some sort of *belief models* in which various aspects of the external reality are encoded. Such belief models provide an explicit or implicit representation for the current state of the world, from the robot's viewpoint. They therefore serve as a representational backbone for a wide range of high-level cognitive capabilities related to reasoning, planning and learning in complex and dynamic environments. They are also essential for the robot to verbalise its own knowledge.

In speech-based HRI, critical tasks in dialogue understanding, management and production are directly dependent on such belief models. Examples are context-sensitive speech recognition [15], reference resolution and generation in small- [11] and large-scale space [24], spoken dialogue parsing [14] and interpretation [20], dialogue management [23], user-tailored response generation [22], and contextually appropriate intonation patterns [13]. Contextual knowledge is also a prerequisite for the dynamic adaptation of the robot's behaviour to different environments and interlocutors [3].

Belief models are usually expressed as high level symbolic representations merging and abstracting information over multiple modalities. For HRI, the incorporated knowledge might include (inter alia): entities in the visual scene, spatial structure, user profiles (intentional and attentional state, preferences), dialogue histories, and task models (what is to be done, which actions are available).

The construction of such belief models raises two important issues. The first question to address is how these high-level representations can be reliably abstracted from low-level sensory data [1], [18]. To be meaningful, most symbolic representations should be *grounded* in (subsymbolic) sensory inputs [19]. This is a difficult problem, partly because of the noise and uncertainty contained in sensory data (partial observability), and partly because the connection between low-level perception and high-level symbols is typically difficult to formalise in a general way [6].

The second issue relates to how information arising from different modalities and time points can be efficiently *merged* into unified multi-modal structures [12], and how these inputs can refine and constrain each other to yield improved estimations, over time. This is the well-known engineering problem of multi-target, multi-sensor data fusion [5].

Belief models are thus the final product of an iterative process of information *fusion*, *refinement* and *abstraction*. Typical HRI environments are challenging to model, being simultaneously *complex*, *multi-agent*, *dynamic* and *uncertain*. Four requirements can be formulated:

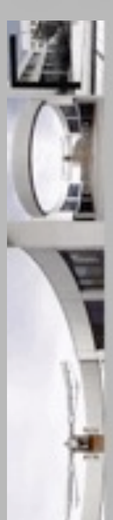
1) HRI environments are complex and reveal a large amount of internal structure (for instance, spatial relations between entities, or groupings of objects). The formal representations used to model them must therefore possess the expressive power to reflect this rich relational structure.
2) Interactive robots are made for multi-agent settings. Making sense of communicative acts requires the ability to distinguish between one's own knowledge (what I believe), knowledge attributed to others (what I think the others believe), and shared common ground knowledge (what we believe as a group).
3) Situated interactions are *dynamic* and evolve over time. The incorporation of spatio-temporal framing is thus necessary to go beyond the "here-and-now" and be capable of linking the present with (episodic) memories of the past and anticipation of future events.
4) And last but not least, due to the partial observability of most contextual features, it is crucial that belief models incorporate an explicit account of *uncertainties*.

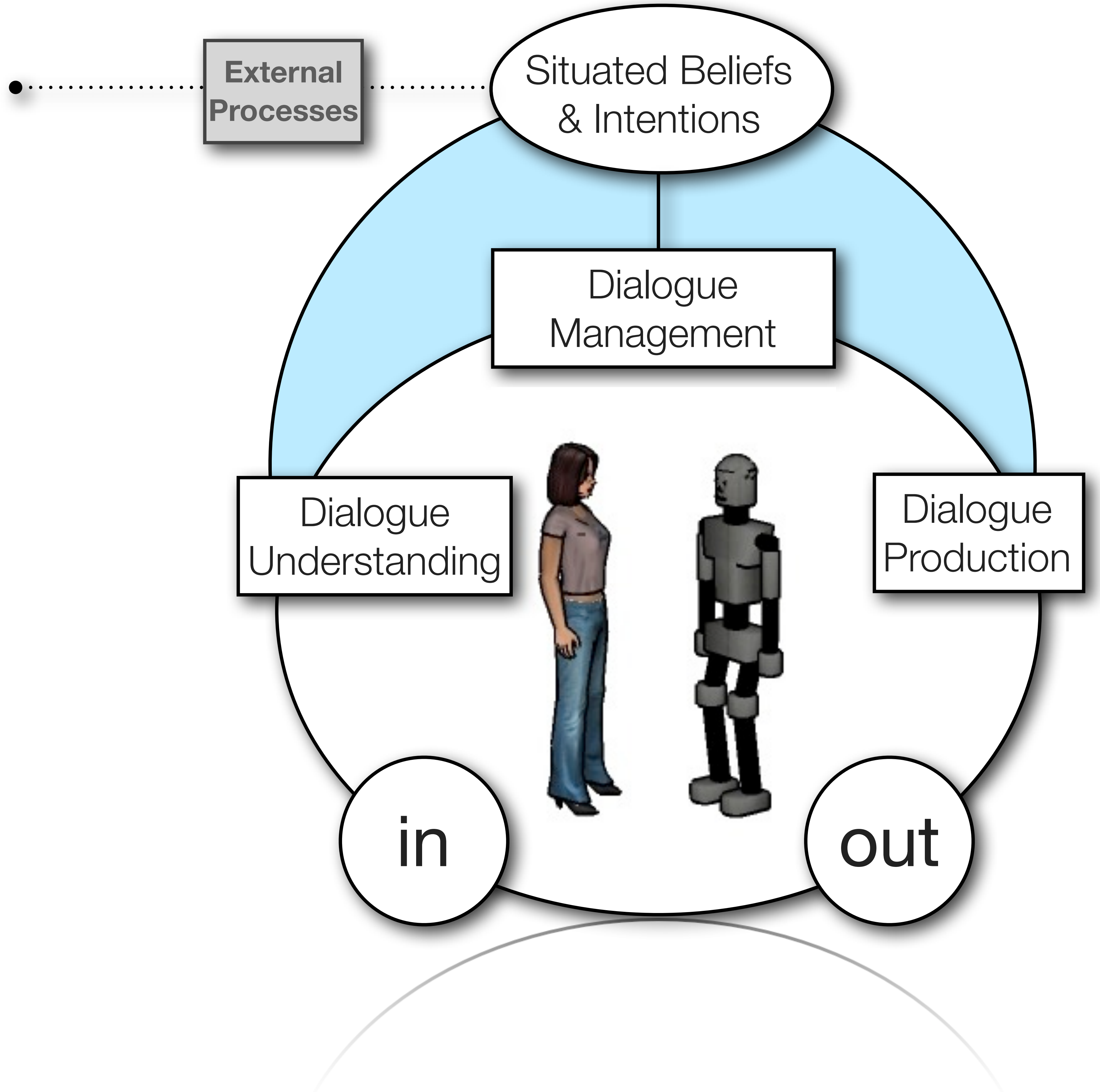
Orthogonal to these "representational" requirements, crucial performance requirements must also be adressed. To keep up with a continuously changing environment, all operations on belief models (updates, queries, etc.) must be performed under soft real-time constraints.

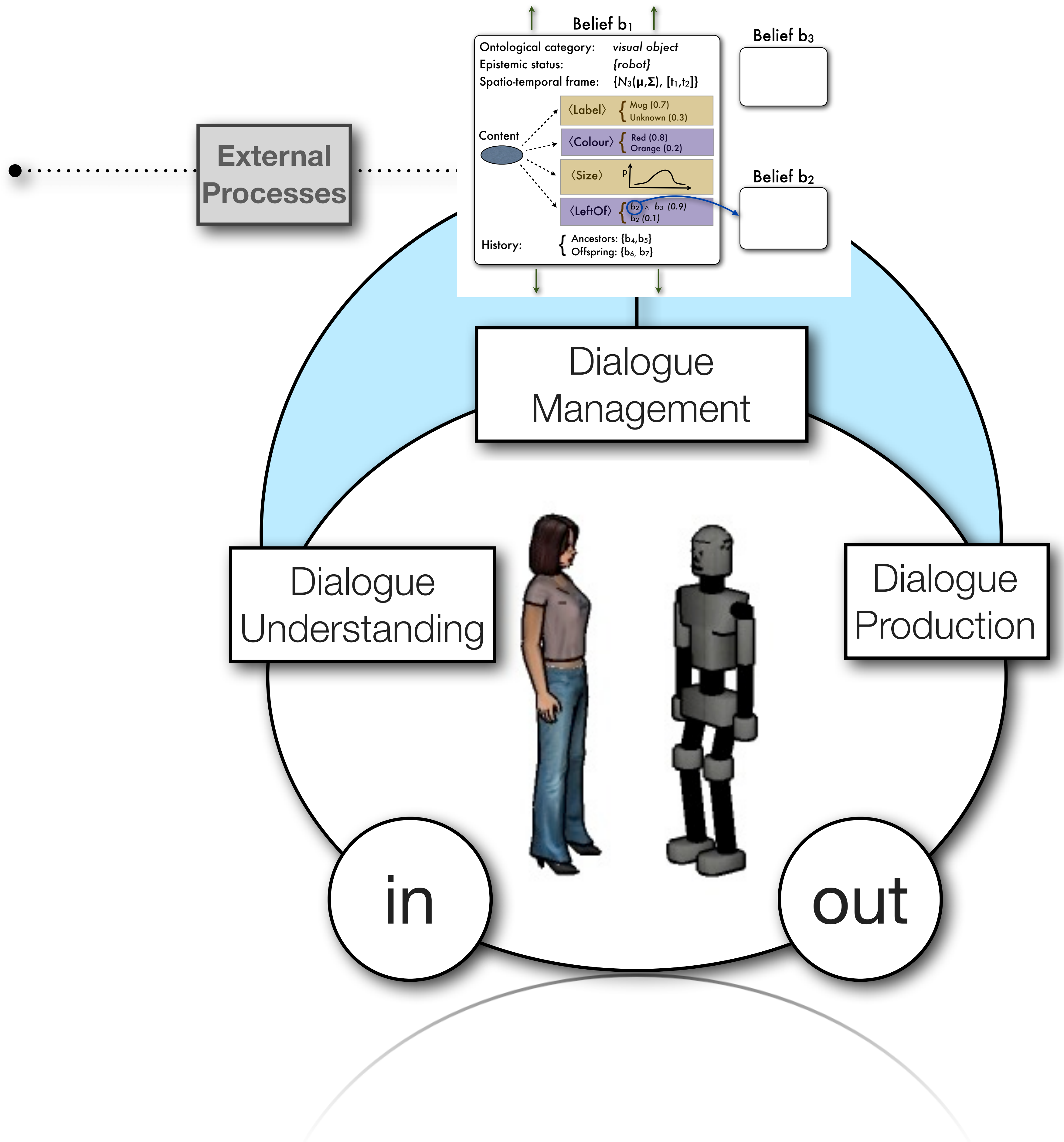
This work was supported by the EU FP7 ICT Integrated Project "*CogX: cognitive systems that self-understand and self-extend*" (FP7-ICT- 215181).

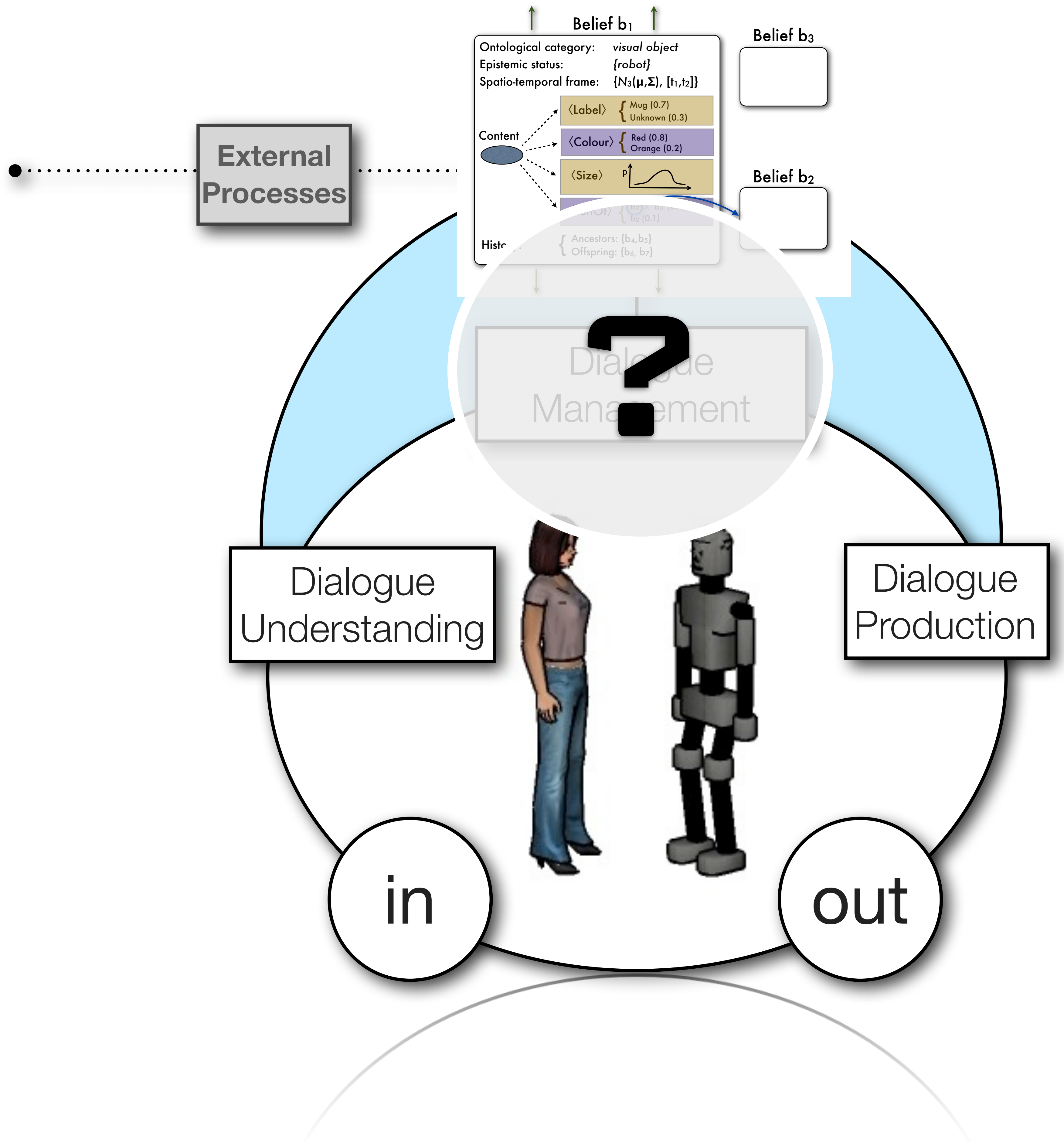
Pierre Lison, Carsten Ehrler and Geert-Jan M. Kruijff are with the German Research Centre for Artificial Intelligence (DFKI GmbH), Language Technology Lab, Saarbrücken, Germany. {plison,carsten.ehrler,gj}@dfki.de

Talking Robots
Language Technology, DFKI




External
Processes
Situated Beliefs
& Intentions
Dialogue
Management
Dialogue
Understanding
Dialogue
Production
in
out


External Processes
Belief b_1
Ontological category: *visual object*
Epistemic status: *{robot}*
Spatio-temporal frame: $\{N_3(\mu,\Sigma), [t_1,t_2]\}$
Content
⟨Label⟩ Mug (0.7) Unknown (0.3)
⟨Colour⟩ Red (0.8) Orange (0.2)
⟨Size⟩ P
⟨LeftOf⟩ $b_2 \wedge b_3$ (0.9) b_2 (0.1)
History: Ancestors: $\{b_4,b_5\}$ Offspring: $\{b_6, b_7\}$
Belief b_3
Belief b_2
Dialogue Management
Dialogue Understanding
Dialogue Production
in
out



Belief b_1
Ontological category: visual object
Epistemic status: {robot}
Spatio-temporal frame: $\{N_3(\mu,\Sigma), [t_1,t_2]\}$
Content
⟨Label⟩ Mug (0.7) Unknown (0.3)
⟨Colour⟩ Red (0.8) Orange (0.2)
⟨Size⟩ P
Ancestors: $\{b_4,b_5\}$
Offspring: $\{b_6, b_7\}$
Belief b_3
Belief b_2
External Processes
Dialogue Management
?
Dialogue Understanding
Dialogue Production
in
out

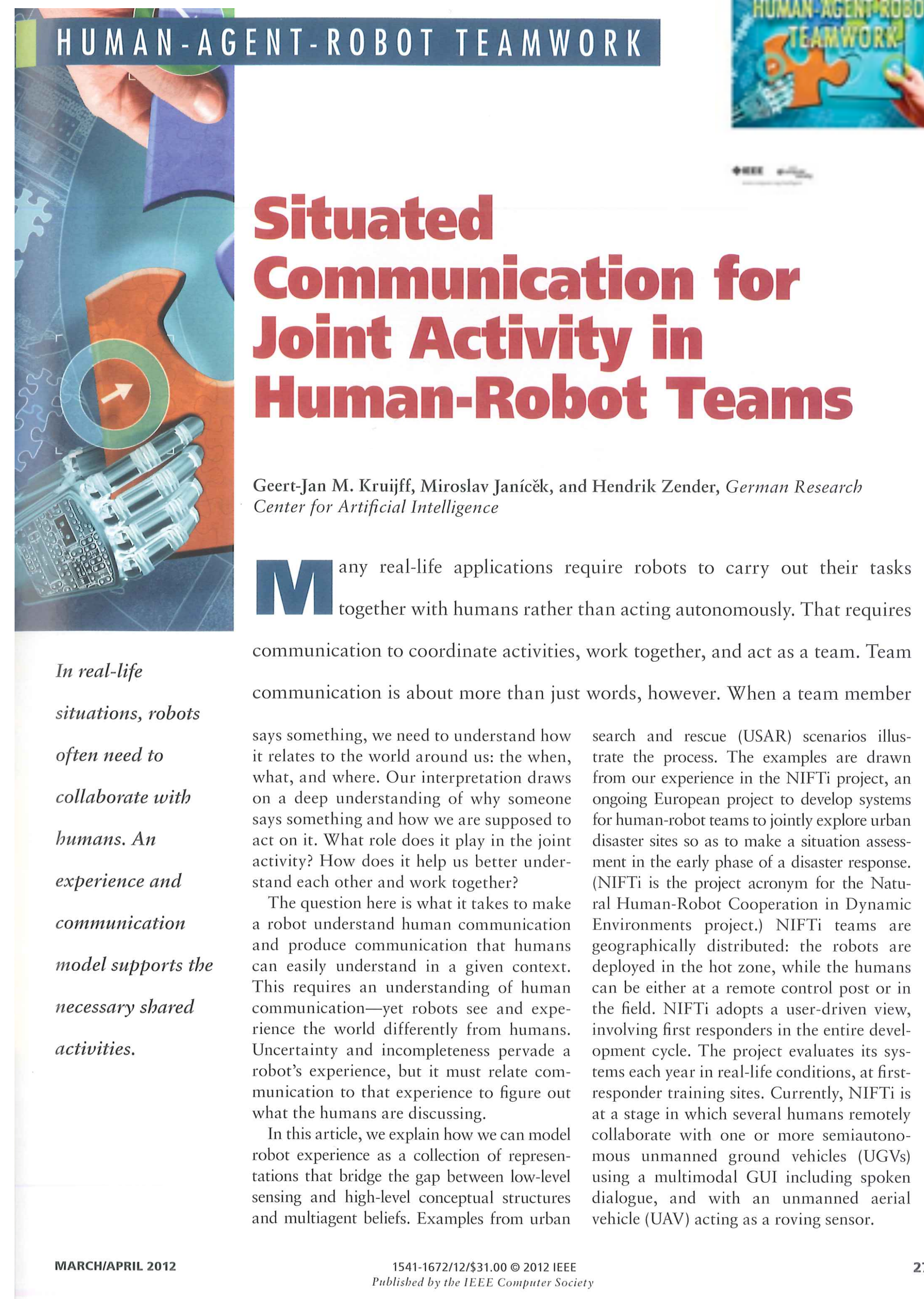

WE CAN DO IT!
©rebecca roka 2011
©Rebecca Roka 2011
RUR
ROBOTS: WORKING SO YOU DON'T HAVE TO.

Collaborative dialogue & Intention

- Problem setting
  - Dialogue is set in, part of, a much wider collaborative activity
  - How can a robot understand intent in that context?
- Development
  - How do intentional context, intension, and extension interact?
  - How can a robot reason about intentions given uncertainty, incomplete knowledge?

Collaborative dialogue & Intention

- Problem setting
  - Dialogue is set in, part of, a much wider collaborative activity
  - How can a robot understand intent in that context?
- Development
  - How do intentional context, intension, and extension interact?
  - How can a robot reason about intentions given uncertainty, incomplete knowledge?



HUMAN-AGENT-ROBOT TEAMWORK

Situated Communication for Joint Activity in Human-Robot Teams

Geert-Jan M. Kruijff, Miroslav Janíček, and Hendrik Zender, *German Research Center for Artificial Intelligence*

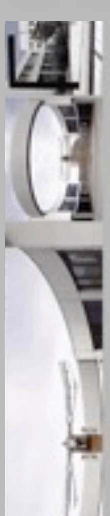
In real-life situations, robots often need to collaborate with humans. An experience and communication model supports the necessary shared activities.

Many real-life applications require robots to carry out their tasks together with humans rather than acting autonomously. That requires communication to coordinate activities, work together, and act as a team. Team communication is about more than just words, however. When a team member says something, we need to understand how it relates to the world around us: the when, what, and where. Our interpretation draws on a deep understanding of why someone says something and how we are supposed to act on it. What role does it play in the joint activity? How does it help us better understand each other and work together?

The question here is what it takes to make a robot understand human communication and produce communication that humans can easily understand in a given context. This requires an understanding of human communication—yet robots see and experience the world differently from humans. Uncertainty and incompleteness pervade a robot's experience, but it must relate communication to that experience to figure out what the humans are discussing.

In this article, we explain how we can model robot experience as a collection of representations that bridge the gap between low-level sensing and high-level conceptual structures and multiagent beliefs. Examples from urban search and rescue (USAR) scenarios illustrate the process. The examples are drawn from our experience in the NIFTi project, an ongoing European project to develop systems for human-robot teams to jointly explore urban disaster sites so as to make a situation assessment in the early phase of a disaster response. (NIFTi is the project acronym for the Natural Human-Robot Cooperation in Dynamic Environments project.) NIFTi teams are geographically distributed: the robots are deployed in the hot zone, while the humans can be either at a remote control post or in the field. NIFTi adopts a user-driven view, involving first responders in the entire development cycle. The project evaluates its systems each year in real-life conditions, at first-responder training sites. Currently, NIFTi is at a stage in which several humans remotely collaborate with one or more semiautonomous unmanned ground vehicles (UGVs) using a multimodal GUI including spoken dialogue, and with an unmanned aerial vehicle (UAV) acting as a roving sensor.

1541-1672/12/$31.00 © 2012 IEEE
Published by the IEEE Computer Society



Abduction

- Intention as explanation
  - Explanation for behavior,
  - *why* something is (to be) said, and *how* the communicated content "fits in" (updates beliefs)
- Abductive inference
  - Given observations *O*, if *P* were the case, then *O;* Hence, reason to believe that *P*
  - Non-monotonicity: *P,* and *P* → *O*
  - *"*Weighted" abduction reformulated as logical-probabilistic reasoning over *assumptions* and *assertions*

IEEE RO-MAN 2010
19th International Symposium on Robot and Human Interactive Communication
September 12-15, 2010 | Principe di Piemonte, Viareggio, Italy

Continual Processing of Situated Dialogue in Human-Robot Collaborative Activities

Geert-Jan M. Kruijff, Miroslav Janíček and Pierre Lison
Language Technology Lab
German Research Center for Artificial Intelligence, DFKI GmbH
{gj,miroslav.janicek,pierre.lison}@dfki.de

Abstract— This paper presents an implemented approach of processing situated dialogue between a human and a robot. The focus is on task-oriented dialogue, set in the larger context of human-robot collaborative activity. The approach models understanding and production of dialogue to include intension (what is being talked about), intention (the goal of why something is being said), and attention (what is being focused on). These dimensions are directly construed in terms of assumptions and assertions on situated multi-agent belief models. The approach is continual in that it allows for interpretations to be dynamically retracted, revised, or deferred. This makes it possible to deal with the inherent asymmetry in how robots and humans tend to understand dialogue, and the world in which it is set. The approach has been fully implemented, and integrated into a cognitive robot. The paper discusses the implementation, and illustrates it in a collaborative learning setting.

I. Introduction

Particularly in task-oriented dialogues between a human and a robot, there is more to dialogue than just understanding words. The robot needs to understand what is being talked about, but it also needs to understand why it was told something. In other words, what the human intends the robot to do with the information, in the larger context of their joint activity.

In this paper we see task-oriented dialogue as part of a larger collaborative activity, in which a human and the robot are involved. They are planning together, executing their plans. Dialogue plays a facilitatory role in this. It helps all participants build up a common ground, and maintain it as plans are executed, and the world around them changes.

We present here an approach that models these aspects of situated task-oriented dialogue. We provide an algorithm in which dialogue is understood, and generated, by looking at *why* something is being said (intention), *what* that something is about (intension), and *how* that helps to direct the focus (attention). Core to the algorithm is abductive reasoning. This type of reasoning tries to find the best explanation for observations. In our case, it tries to find the best explanation for why something was said (understanding), or how an intention best could be achieved communicatively (generation). Thereby, abduction directly works off the situated, multi-agent belief models the robot maintains as part of its understanding of the world, and of the agents acting therein.

Our approach views dialogue from a more intentional perspective, like the work by Grosz & Sidner [6], Lochbaum et al. [10], and most recently Stone et al [14], [15], [16]. Our approach extends that of Stone et al.

Stone et al. formulate an algorithm for collaborative activity, involving abductive reasoning. They assume that understanding and production are symmetric: "what I say is how you understand it". However, this is optimistic for human-human dialogue, and rather unrealistic for human-robot interaction. Robots hardly ever perfectly understand what is meant. We need to allow for the robot to act upon interpretations even when they are incomplete or uncertain. And, should it turn out that the robot has misunderstood what was said, roll dialogue back to a point where the robot can clarify and correct its understanding.

Our approach enables these features by introducing *assertions* into our logics. This idea is inspired by Brenner & Nebel's work on continual planning [3]. An assertion is a content formula that needs to be verified at a later point. In that, it is different from a propositional fact, which the robot knows to be either true or false. We can introduce an assertion into an abductive inference to help find an explanation, and then act upon it. It is just that this is then made contingent on the assertion to become true sooner or later. In this paper, we show how assertions can play a fundamental role in helping a robot and a human achieve common ground in collaborative activity.

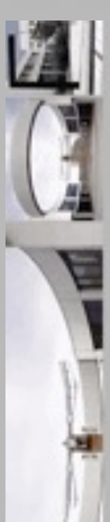
Below, §II provides a brief overview on intentional approaches to dialogue. §III presents our approach and discusses situated multi-agent belief models, abductive reasoning, and the algorithm for continual processing of collaborative activity. §IV discusses the implementation, and §V illustrates it on working examples from an integrated robot system.

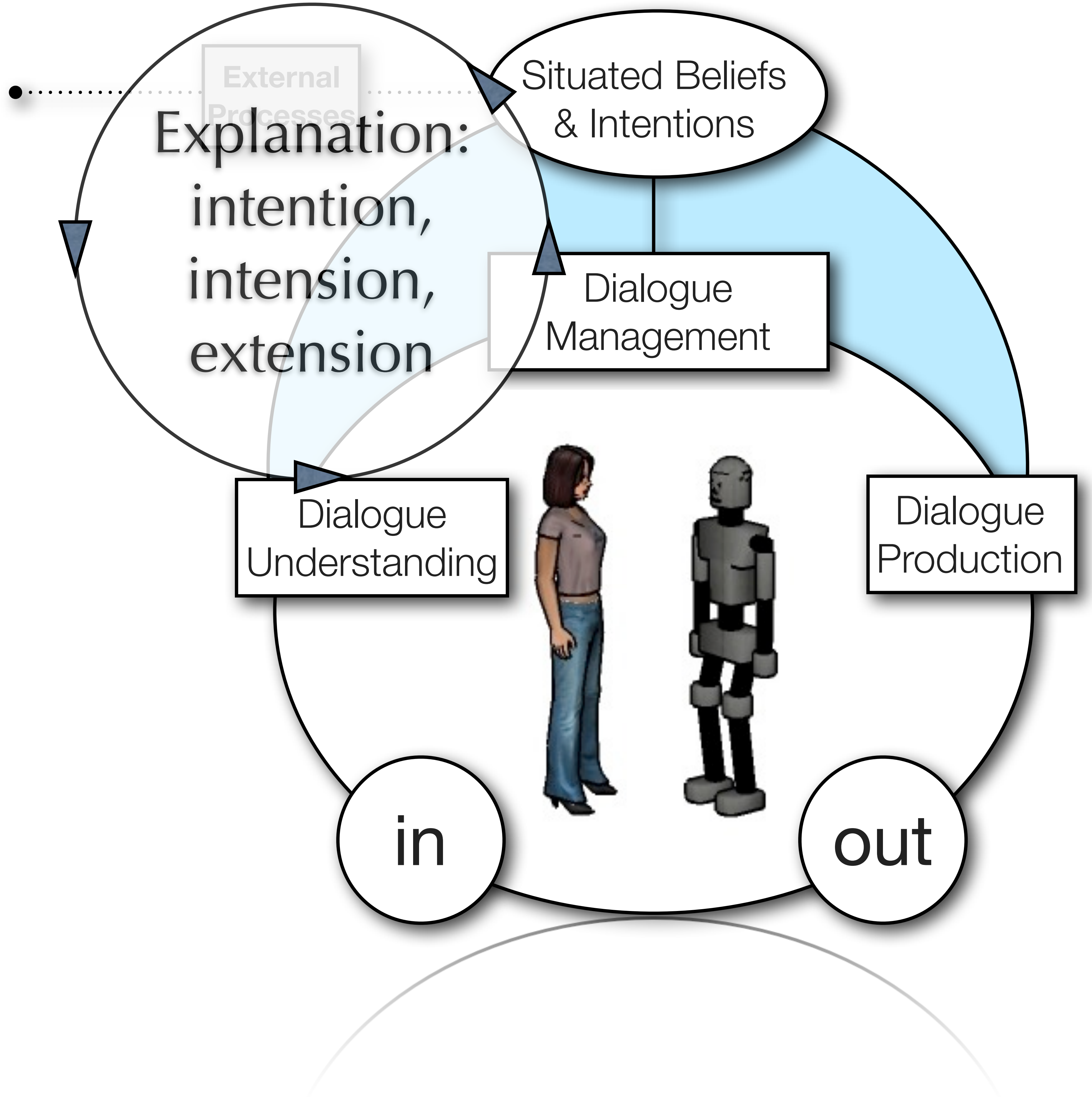
II. Background

Recent theories of dialogue focus on how participants can obtain common ground through alignment [11]. Agents align how they communicate content, what they pay attention to, and what they intend to do next. They base this on how they perceive each other's views on the world.

This works out reasonably well as long as we can assume a more or less common way of "looking" at things. Even when humans normally differ in what they know, can, and intend to do, there is typically a common categorical framework in which they can characterize the world, in order to arrive at a

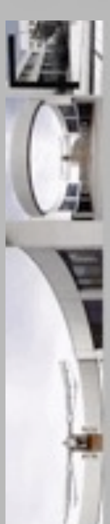
Talking Robots
Language Technology, DFKI




External
Processes
Explanation:
intention,
intension,
extension
Situated Beliefs
& Intentions
Dialogue
Management
Dialogue
Understanding
Dialogue
Production
in
out

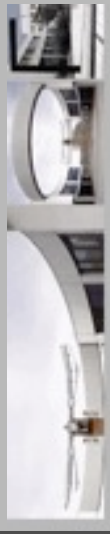
Example: Functional mapping

- Intention recognition
  - Intention-in-referential-context
  - Explanation = Belief about what to do, where, and why
- Example
  - Functional mapping: Projection of affordances into situation
  - “Go to the car” = Command to go to a particular area next to a car from where it is expectedly optimally possible for the robot to look inside said car to see whether there are victims inside



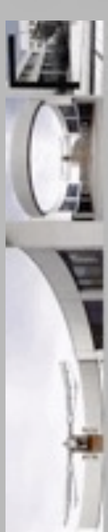
Example: Functional mapping

- Intention recognition
  - Intention-in-referential-context
  - Explanation = Belief about what to do, where, and why
- Example
  - Functional mapping: Projection of affordances into situation
  - "Go to the car" = Command to go to a particular area next to a car from where it is expectedly optimally possible for the robot to look inside said car to see whether there are victims inside



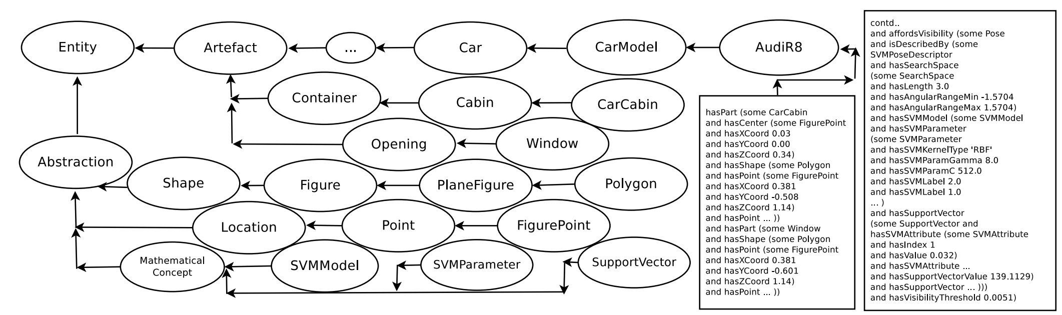
Example: Functional mapping

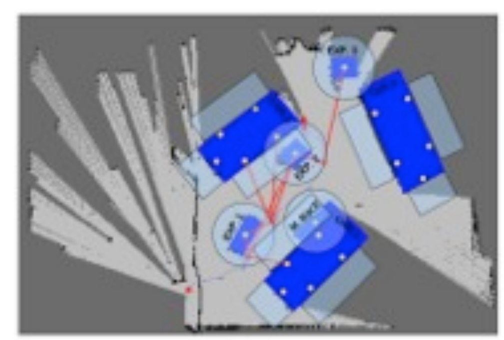
- Intention recognition
  - Intention-in-referential-context
  - Explanation = Belief about what to do, where, and why
- Example
  - Functional mapping: Projection of affordances into situation
  - “Go to the car” = Command to go to a particular area next to a car from where it is expectedly optimally possible for the robot to look inside said car to see whether there are victims inside

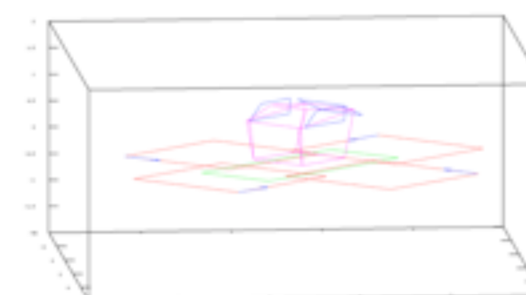


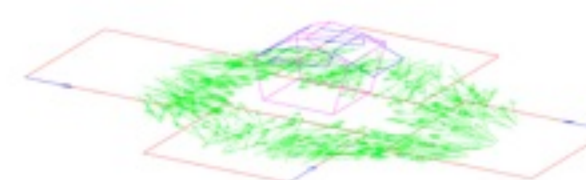
Example: Functional mapping

- Intention recognition
  - Intention-in-referential-context
  - Explanation = Belief about what to do, where, and why
- Example
  - Functional mapping: Projection of affordances into situation
  - “Go to the car” = Command to go to a particular area next to a car from where it is expectedly optimally possible for the robot to look inside said car to see whether there are victims inside



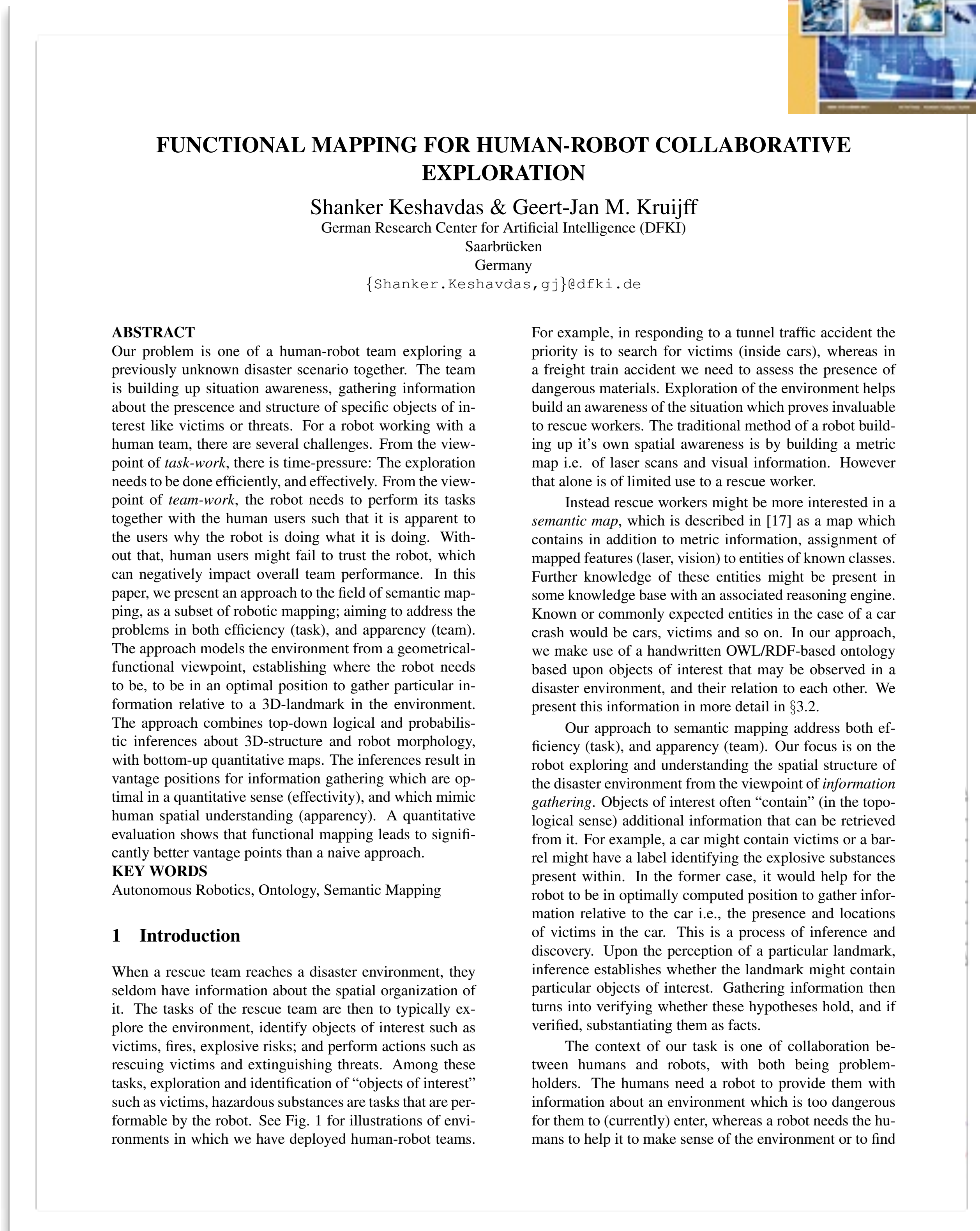






Example: Functional mapping

- Intention recognition
  - Intention-in-referential-context
  - Explanation = Belief about what to do, where, and why
- Example
  - Functional mapping: Projection of affordances into situation
  - "Go to the car" = Command to go to a particular area next to a car from where it is expectedly optimally possible for the robot to look inside said car to see whether there are victims inside



FUNCTIONAL MAPPING FOR HUMAN-ROBOT COLLABORATIVE EXPLORATION

Shanker Keshavdas & Geert-Jan M. Kruijff
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken
Germany
{Shanker.Keshavdas,gj}@dfki.de

ABSTRACT
Our problem is one of a human-robot team exploring a previously unknown disaster scenario together. The team is building up situation awareness, gathering information about the prescence and structure of specific objects of interest like victims or threats. For a robot working with a human team, there are several challenges. From the viewpoint of *task-work*, there is time-pressure: The exploration needs to be done efficiently, and effectively. From the viewpoint of *team-work*, the robot needs to perform its tasks together with the human users such that it is apparent to the users why the robot is doing what it is doing. Without that, human users might fail to trust the robot, which can negatively impact overall team performance. In this paper, we present an approach to the field of semantic mapping, as a subset of robotic mapping; aiming to address the problems in both efficiency (task), and apparency (team). The approach models the environment from a geometrical-functional viewpoint, establishing where the robot needs to be, to be in an optimal position to gather particular information relative to a 3D-landmark in the environment. The approach combines top-down logical and probabilistic inferences about 3D-structure and robot morphology, with bottom-up quantitative maps. The inferences result in vantage positions for information gathering which are optimal in a quantitative sense (effectivity), and which mimic human spatial understanding (apparency). A quantitative evaluation shows that functional mapping leads to significantly better vantage points than a naive approach.

KEY WORDS
Autonomous Robotics, Ontology, Semantic Mapping

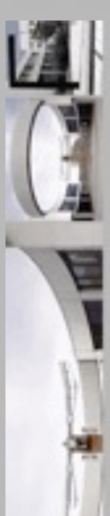
1 Introduction

When a rescue team reaches a disaster environment, they seldom have information about the spatial organization of it. The tasks of the rescue team are then to typically explore the environment, identify objects of interest such as victims, fires, explosive risks; and perform actions such as rescuing victims and extinguishing threats. Among these tasks, exploration and identification of "objects of interest" such as victims, hazardous substances are tasks that are performable by the robot. See Fig. 1 for illustrations of environments in which we have deployed human-robot teams. For example, in responding to a tunnel traffic accident the priority is to search for victims (inside cars), whereas in a freight train accident we need to assess the presence of dangerous materials. Exploration of the environment helps build an awareness of the situation which proves invaluable to rescue workers. The traditional method of a robot building up it's own spatial awareness is by building a metric map i.e. of laser scans and visual information. However that alone is of limited use to a rescue worker.

Instead rescue workers might be more interested in a *semantic map*, which is described in [17] as a map which contains in addition to metric information, assignment of mapped features (laser, vision) to entities of known classes. Further knowledge of these entities might be present in some knowledge base with an associated reasoning engine. Known or commonly expected entities in the case of a car crash would be cars, victims and so on. In our approach, we make use of a handwritten OWL/RDF-based ontology based upon objects of interest that may be observed in a disaster environment, and their relation to each other. We present this information in more detail in §3.2.

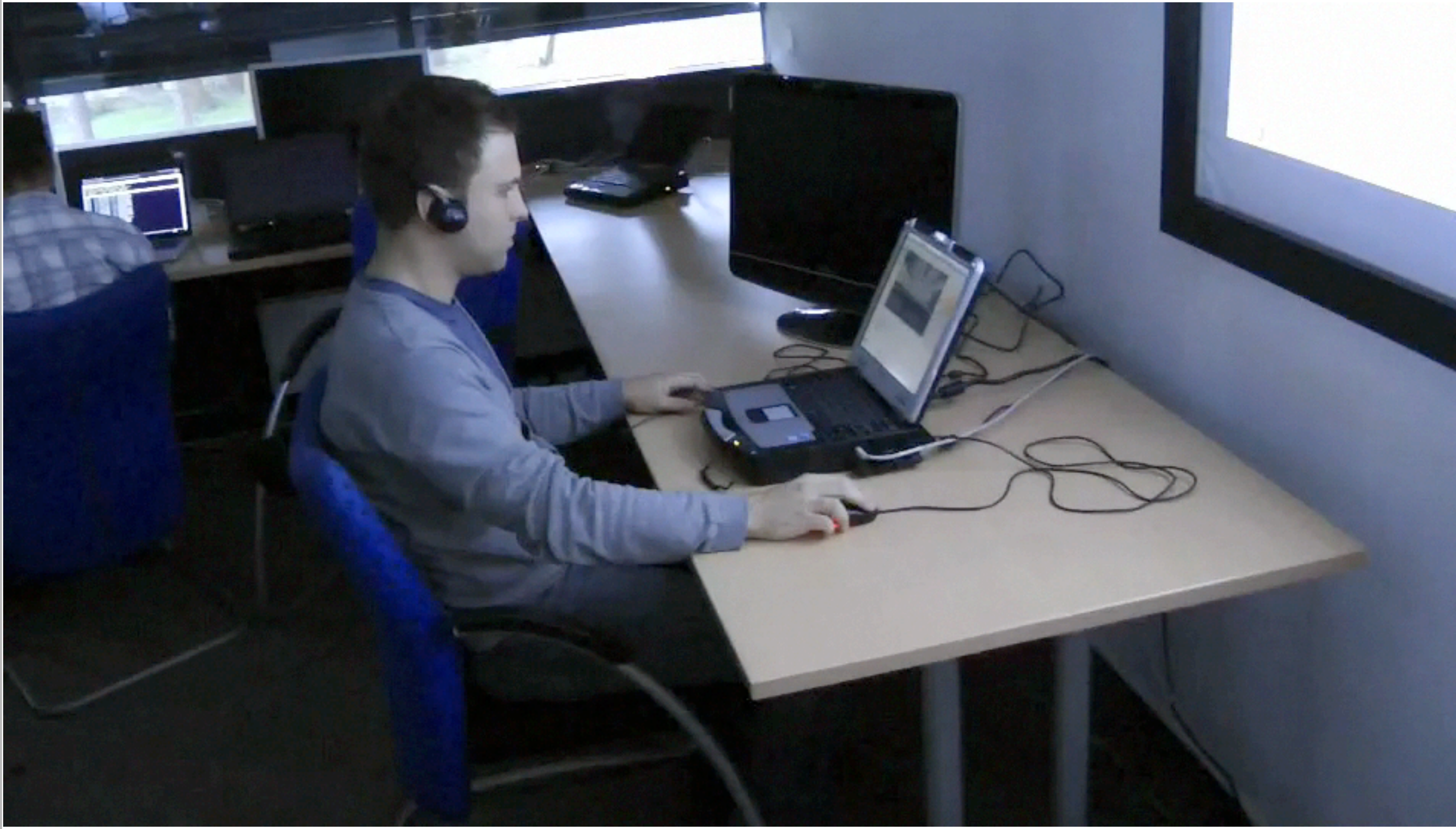
Our approach to semantic mapping address both efficiency (task), and apparency (team). Our focus is on the robot exploring and understanding the spatial structure of the disaster environment from the viewpoint of *information gathering*. Objects of interest often "contain" (in the topological sense) additional information that can be retrieved from it. For example, a car might contain victims or a barrel might have a label identifying the explosive substances present within. In the former case, it would help for the robot to be in optimally computed position to gather information relative to the car i.e., the presence and locations of victims in the car. This is a process of inference and discovery. Upon the perception of a particular landmark, inference establishes whether the landmark might contain particular objects of interest. Gathering information then turns into verifying whether these hypotheses hold, and if verified, substantiating them as facts.

The context of our task is one of collaboration between humans and robots, with both being problem-holders. The humans need a robot to provide them with information about an environment which is too dangerous for them to (currently) enter, whereas a robot needs the humans to help it to make sense of the environment or to find

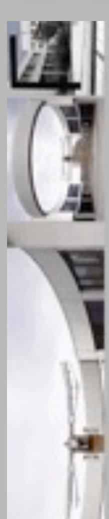


Example: instruction & mapping (2011)

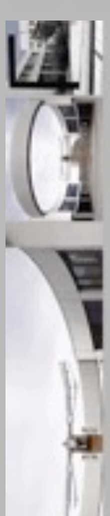
Example: instruction & mapping (2011)



Example: functional mapping (2011)

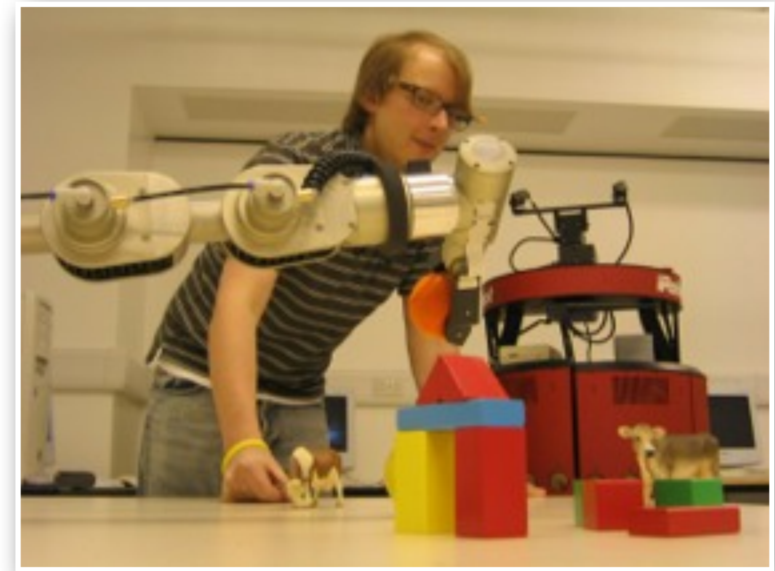


Example: functional mapping (2011)



Using Task-Context

- Priming using situation + task
  - Understanding is context-sensitive
  - Use salient situation (objects, features), task context, to prime
- Priming in ASR
  - Lexical activation network to bias Language Model used in ASR
  - Dialogue state, situation, tasks
  - -16.1% reduction in WER over baseline (Nuance v8.5)
- Priming in parsing
  - Parsing with CCGs: grammatical **and non-grammatical** constructions to deal with complexity of dialogue
  - Discriminative parse selection during incremental parsing with CCG, needed to deal with #analyses
  - Use of audio, syntactic, semantic and context features to train perceptrons for parse selection models
  - Further reduction in WER (^semantics), Δ=-23.4%
  - Parsing time reduced by Δ=-51.9% without significant loss in performance against Gold Standard



EXAMPLE

- Given a visually salient red block
- Recognized by the robot as such
- Lexical association connects this “red block” to words like “block”, “square”, “pick up”, etc.
- The language model is adapted to increase the probability of hearing these words.

Using Task-Context

- Priming using situation + task
  - Understanding is context-sensitive
  - Use salient situation (objects, features), task context, to prime
- Priming in ASR
  - Lexical activation network to bias Language Model used in ASR
  - Dialogue state, situation, tasks
  - -16.1% reduction in WER over baseline (Nuance v8.5)
- Priming in parsing
  - Parsing with CCGs: grammatical **and non-grammatical** constructions to deal with complexity of dialogue
  - Discriminative parse selection during incremental parsing with CCG, needed to deal with #analyses
  - Use of audio, syntactic, semantic and context features to train perceptrons for parse selection models
  - Further reduction in WER (^semantics), Δ=-23.4%
  - Parsing time reduced by Δ=-51.9% without significant loss in performance against Gold Standard

Salience-driven Contextual Priming of Speech Recognition for Human-Robot Interaction

Pierre Lison and Geert-Jan Kruijff [1]

Abstract. The paper presents an implemented model for priming speech recognition, using contextual information about salient entities. The underlying hypothesis is that, in human-robot interaction, speech recognition performance can be improved by exploiting knowledge about the immediate physical situation and the dialogue history. To this end, visual salience (objects perceived in the physical scene) and linguistic salience (objects, events already mentioned in the dialogue) are integrated into a single cross-modal salience model. The model is dynamically updated as the environment changes. It is used to establish expectations about which words are most likely to be heard in the given context. The update is realised by continuously adapting the word-class probabilities specified in a statistical language model. The paper discusses the motivations behind the approach, and presents the implementation as part of a cognitive architecture for mobile robots. Evaluation results on a test suite show a statistically significant improvement of salience-driven priming speech recognition (WER) over a commercial baseline system.

1 Introduction

Service robots are becoming more and more sophisticated. In many cases, these robots must operate in open-ended environments and interact with humans using spoken natural language to perform a variety of service-oriented tasks. This has led to an increasing interest in developing dialogue systems for robots [28, 15, 23]. A fundamental challenge here is, how the robot can *situate* the dialogue: The robot should be able to understand what is being said, *and* how that relates to the physical situation [20, 25, 26, 11].

The relation between language and experience is often characterized as being *bi-directional* (cf. [14]). That is, language influences how to perceive the environment – and vice versa, the physical situation provides a context against which to interpret language. In this paper, we focus on how information from the dialogue- and situated context can help guiding, and improving, automatic speech recognition (ASR) in human-robot interaction (HRI). Spoken dialogue is one of the most natural means of communication for humans. Despite significant technological advances, however, ASR remains for most tasks at least an order of magnitude worse than that of human listeners [17]. This particularly holds for using ASR in HRI systems which typically have to operate in real-world noisy environments, dealing with utterances pertaining to complex, open-ended domains.

In this paper we present an approach to using context in priming ASR. By priming we mean, focusing the domain of words / word sequences ASR can expect next, so as to improve recognition. This approach has been implemented, and integrated into a cognitive architecture for a mobile robot [10, 14]. Evaluation results on a test suite with recordings of "free speech" in the application domain show a statistically significant decrease in word-error rate (WER) of the implemented system, over a commercial baseline system.

We follow [9] and use context information (in the form of contextual constraints) to update the statistical language model used in ASR. We define a *context-sensitive language model* which exploits information about salient objects in the visual scene and linguistic expressions in the dialogue history to prime recognition. A *salience model* integrating both visual and linguistic salience [12] is used to dynamically compute lexical activations, which are incorporated into the language model at runtime.

The structure of the paper is as follows. We first situate our approach against the background of situated dialogue and ASR, and introduce the software architecture in which our system has been integrated. We then describe the salience model, and explain how it is utilised within the language model used for ASR. We finally present the evaluation of our approach, followed by conclusions.

Figure 1. Example interaction

2 Background

The combinatorial nature of language provides virtually unlimited ways in which we can communicate meaning. This, of course, raises the question of how precisely an utterance should then be understood as it is being heard. Empirical studies have investigated what information humans use when comprehending spoken utterances. An important observation is that interpretation *in context* plays a crucial role in the comprehension of utterance as it unfolds [13]. During utterance comprehension, humans combine linguistic information with scene understanding and "world knowledge."

Several approaches in processing situated dialogue for HRI have made similar observations [19, 20, 21, 4, 14]: A robot's understanding can be improved by relating utterances to the situated context. This first of all presumes the robot *is able to* relate language and the world around it. [22] present a comprehensive overview of existing

[1] DFKI GmbH, Saarbrücken, Germany, email: {pierre.lison} {gj}@dfki.de

Talking Robots
Language Technology, DFKI

Using Task-Context

- Priming using situation + task
  - Understanding is context-sensitive
  - Use salient situation (objects, features), task context, to prime
- Priming in ASR
  - Lexical activation network to bias Language Model used in ASR
  - Dialogue state, situation, tasks
  - -16.1% reduction in WER over baseline (Nuance v8.5)
- Priming in parsing
  - Parsing with CCGs: grammatical **and non-grammatical** constructions to deal with complexity of dialogue
  - Discriminative parse selection during incremental parsing with CCG, needed to deal with #analyses
  - Use of audio, syntactic, semantic and context features to train perceptrons for parse selection models
  - Further reduction in WER (^semantics), Δ=-23.4%
  - Parsing time reduced by Δ=-51.9% without significant loss in performance against Gold Standard

ECAI Patras 2008

IEEE RO-MAN 2009

Efficient Parsing of Spoken Inputs for Human-Robot Interaction

Pierre Lison and Geert-Jan M. Kruijff

Abstract— **The use of deep parsers in spoken dialogue systems is usually subject to strong performance requirements. This is particularly the case in human-robot interaction, where the computing resources are limited and must be shared by many components in parallel. A real-time dialogue system must be capable of responding quickly to any given utterance, even in the presence of noisy, ambiguous or distorted input. The parser must therefore ensure that the number of analyses remains bounded at every processing step.**

The paper presents a practical approach to addressing this issue in the context of deep parsers designed for spoken dialogue. The approach is based on a word lattice parser combined with a statistical model for parse selection. Each word lattice is parsed incrementally, word by word, and a discriminative model is applied at each incremental step to prune the set of resulting partial analyses. The model incorporates a wide range of linguistic and contextual features and can be trained with a simple perceptron. The approach is fully implemented as part of a spoken dialogue system for human-robot interaction. Evaluation results on a Wizard-of-Oz test suite demonstrate significant improvements in parsing time.

I. Introduction

Most dialogue systems developed nowadays for human-robot interaction are based on crude processing methods such as keyword spotting or heuristic rules. These methods are undoubtedly useful for well-structured tasks definable by a set of slot-value pairs, but do not extend very well to more complex interactions, because they are insensitive to the syntactic and semantic structure of the utterance. To capture these linguistic relations, we need to build fine-grained *grammars* of natural language, as well as *parsers* operating on these grammars. Yet, the development of robust and efficient parsers for spoken dialogue is hindered by several major difficulties which need to be addressed.

The first difficulty is the pervasiveness of *speech recognition errors*. Automatic speech recognition is a highly error-prone task, and parsers designed to process spoken input must therefore be able to accomodate the various recognition errors that may arise. This problem is particularly acute for robots operating in real-world environments and dealing with utterances pertaining to complex, open-ended domains.

The second issue is the *relaxed grammaticality* of spoken language. Dialogue utterances are often incomplete, fragmentary or ungrammatical, and may contain numerous disfluencies like fillers (err, uh, mm), repetitions, self-corrections, etc. This is natural behaviour in human-human interaction [1] and can also be observed in several domain-specific corpora for HRI [2]. Spoken dialogue parsers should therefore be made robust to such ill-formed utterances.

Finally, the vast majority of spoken dialogue systems are designed to operate in *real-time*. This has two important consequences. First, the parser should not wait for the utterance to be completed to start processing it – instead, the set of possible semantic interpretations should be gradually built and extended as the utterance unfolds. Second, each processing step should operate under strict time constraints. The main obstacle here is the high level of ambiguity in natural language, which can lead to a combinatorial explosion in the number of possible readings.

The remainder of this paper is devoted to addressing this last issue, building on an integrated approach to situated spoken dialogue processing previously outlined in [3], [4]. The approach we present here is similar to [5], with some notable differences concerning the parser (our parser being specifically tailored for robust spoken dialogue processing), and the features included in the discriminative model.

An overview of the paper is as follows. We first describe in Section II the cognitive architecture in which our system has been integrated. We then discuss the approach in detail in Section III. Finally, we present in Section IV the quantitative evaluations on a WOZ test suite, and conclude.

II. Architecture

The approach we present in this paper is fully implemented and integrated into a cognitive architecture for autonomous robots. A recent description of the architecture is provided in [6], [7]. It is capable of building up visuo-spatial models of a dynamic local scene, and continuously plan and execute manipulation actions on objects within that scene. The robot can discuss objects and their material- and spatial properties for the purpose of visual learning and manipulation tasks. Figure 1 illustrates the architecture schema for the communication subsystem, limited to the comprehension side.

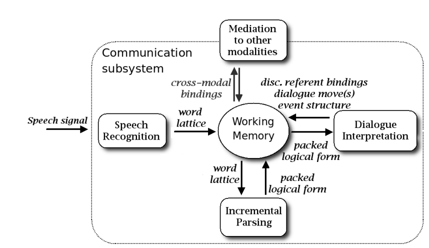
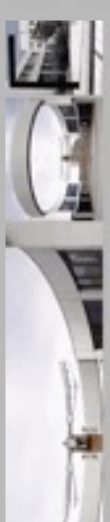
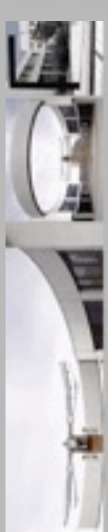


Fig. 1. Architecture schema of the communication subsystem.

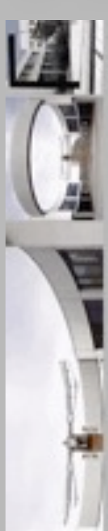
This work was supported by the EU FP7 ICT Integrated Project "CogX" (FP7-ICT- 215181).

Pierre Lison and Geert-Jan M. Kruijff are with the German Research Centre for Artificial Intelligence (DFKI GmbH), Language Technology Lab, Saarbrücken, Germany {pierre.lison,gj} @ dfki.de

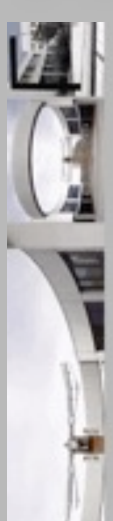




Collaboration



(Enter: Reality)

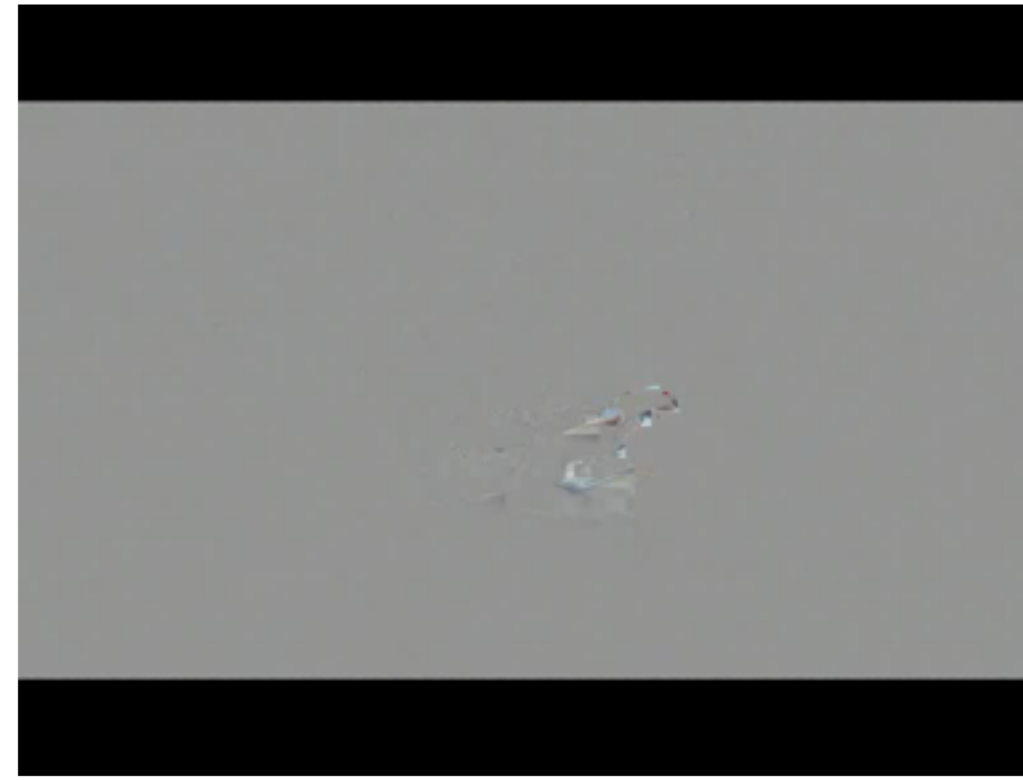


Social context

- Real end users
  - Hospitalized children, pediatricians
  - Rescue workers in disaster response settings
  - Real-life collaborative context
- Impact on dialogue
  - Dialogue as social, situated construction of meaning,
  - set within collaborative activity.
  - Adaptation, dynamics of dialogue reflects dynamics of social context

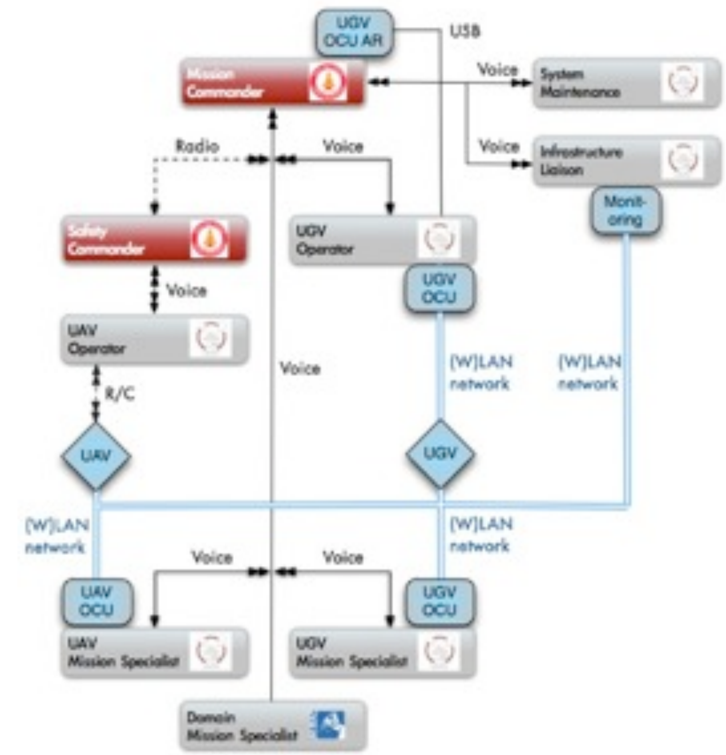
Social context

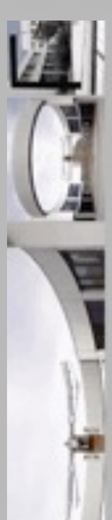
- Real end users
  - Hospitalized children, pediatricians
  - Rescue workers in disaster response settings
  - Real-life collaborative context
- Impact on dialogue
  - Dialogue as social, situated construction of meaning,
  - set within collaborative activity.
  - Adaptation, dynamics of dialogue reflects dynamics of social context



Perspective

- Expression of social dynamics
  - Different roles in a human-robot team
  - Roles imply different perspectives on reality, and on (salient) information
- Subjective nature of meaning
  - Robots and humans perceive reality differently ~ subjectively
  - No "objective truth" but alignment of subjective truths
  - Changes the nature of "shared beliefs" and "common ground"
  - Proof-based judgments & alignment





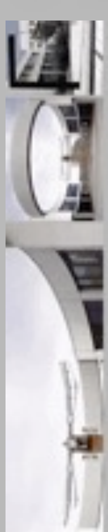
Perspective

- Expression of social dynamics
  - Different roles in a human-robot team
  - Roles imply different perspectives on reality, and on (salient) information
- Subjective nature of meaning
  - Robots and humans perceive reality differently ~ subjectively
  - No "objective truth" but alignment of subjective truths
  - Changes the nature of "shared beliefs" and "common ground"
  - Proof-based judgments & alignment

Perspective

- Expression of social dynamics
  - Different roles in a human-robot team
  - Roles imply different perspectives on reality, and on (salient) information
- Subjective nature of meaning
  - Robots and humans perceive reality differently ~ subjectively
  - No “objective truth” but alignment of subjective truths
  - Changes the nature of “shared beliefs” and “common ground”
  - Proof-based judgments & alignment

UAV Mission Spc:
Robot-Ego-centric

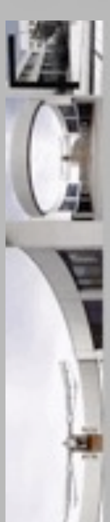


Perspective

- Expression of social dynamics
  - Different roles in a human-robot team
  - Roles imply different perspectives on reality, and on (salient) information
- Subjective nature of meaning
  - Robots and humans perceive reality differently ~ subjectively
  - No "objective truth" but alignment of subjective truths
  - Changes the nature of "shared beliefs" and "common ground"
  - Proof-based judgments & alignment

Perspective

- Expression of social dynamics
  - Different roles in a human-robot team
  - Roles imply different perspectives on reality, and on (salient) information
- Subjective nature of meaning
  - Robots and humans perceive reality differently ~ subjectively
  - No “objective truth” but alignment of subjective truths
  - Changes the nature of “shared beliefs” and “common ground”
  - Proof-based judgments & alignment



Perspective

- Expression of social dynamics
  - Different roles in a human-robot team
  - Roles imply different perspectives on reality, and on (salient) information
- Subjective nature of meaning
  - Robots and humans perceive reality differently ~ subjectively
  - No “objective truth” but alignment of subjective truths
  - Changes the nature of “shared beliefs” and “common ground”
  - Proof-based judgments & alignment

Künstl Intell
DOI 10.1007/s13218-013-0238-3

TECHNICAL CONTRIBUTION

Symbol Grounding as Social, Situated Construction of Meaning in Human-Robot Interaction

Geert-Jan M. Kruijff

Received: 1 October 2012 / Accepted: 8 March 2013

Abstract The paper views the issue of “symbol grounding” from the viewpoint of the construction of meaning *between* humans and robots, in the context of a collaborative activity. This concerns a core aspect of the formation of common ground: The construction of meaning between actors as a conceptual representation which is believed to be mutually understood as referring to a particular aspect of reality. The problem in this construction is that experience is inherently subjective—and more specifically, humans and robots experience and understand reality fundamentally differently. There is an inherent asymmetry between the actors involved. The paper focuses on how this asymmetry can be reflected logically, and particularly in the underlying model theory. The point is to make it possible for a robot to reason explicitly both about such asymmetry in understanding, consider possibilities for alignment to deal with it, and establish (from its viewpoint) a level of intersubjective or mutual understanding. Key to the approach taken in the paper is to consider conceptual representations to be formulas over propositions which are based in *proofs*, as reasoned explanations of experience. This shifts the focus from a notion of “truth” to a notion of judgment—judgments which can be subjectively right and still intersubjectively wrong (faultless disagreement), and which can evolve over time (updates, revision). The result is an approach which accommodates both asymmetric agency and social sentience, modelling symbol grounding in human-robot interaction as social, situated construction over time.

G.-J.M. Kruijff (✉)
Talking Robots@the Language Technology Lab, DFKI GmbH, Saarbrücken, Germany
e-mail: gj@dfki.de

Keywords Human-robot interaction · Multi-agent collaboration · Situation awareness · Common ground

1 Introduction

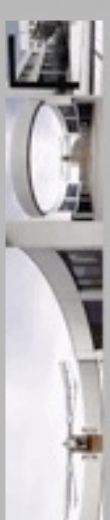
Symbol grounding [32] can be understood as covering a wide range of problems, all regarding the construction of “symbols” or conceptual representations, for understanding experience. For any actor, this starts in forming an understanding of one’s own, subjective experience. But then, when it comes to collaborative activity between actors, we face an additional problem. A “common ground” or level of mutual understanding of reality needs to be created, to collaborate successfully [12, 13]. From the viewpoint of symbol grounding, an individual actor needs to project how another might experience reality, and how that resulting (projected) understanding could be aligned to its own—to then come to a level of intersubjective (aligned) understanding [30, 53].

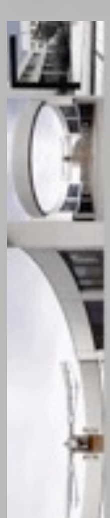
This holds for human-robot interaction just as much as it does for human-human interaction. Humans and robots also need to form a level of common ground to successfully collaborate, and communicate [35, 36, 43, 65]. Without common ground, communication and collaboration typically break down. It becomes difficult for humans to understand why a robot does (or does not) behave the way it does (lack of “transparency” [67] or “apparency”), thus degrading human trust in the system. The kind of situations this can lead to is illustrated e.g. by the loss of a Quince robot during the disaster response at the Fukushima nuclear power plant [29].

What makes the case of human-robot interaction particularly difficult is that there is a fundamental asymmetry

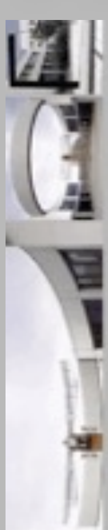
Published online: 28 March 2013 Springer

Talking Robots
Language Technology, DFKI



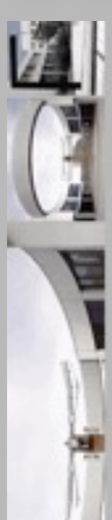


[incr reality]

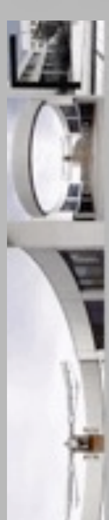


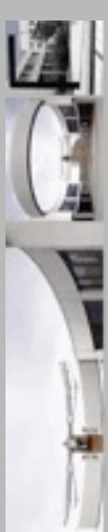
Harsh. Dangerous. Stressful.

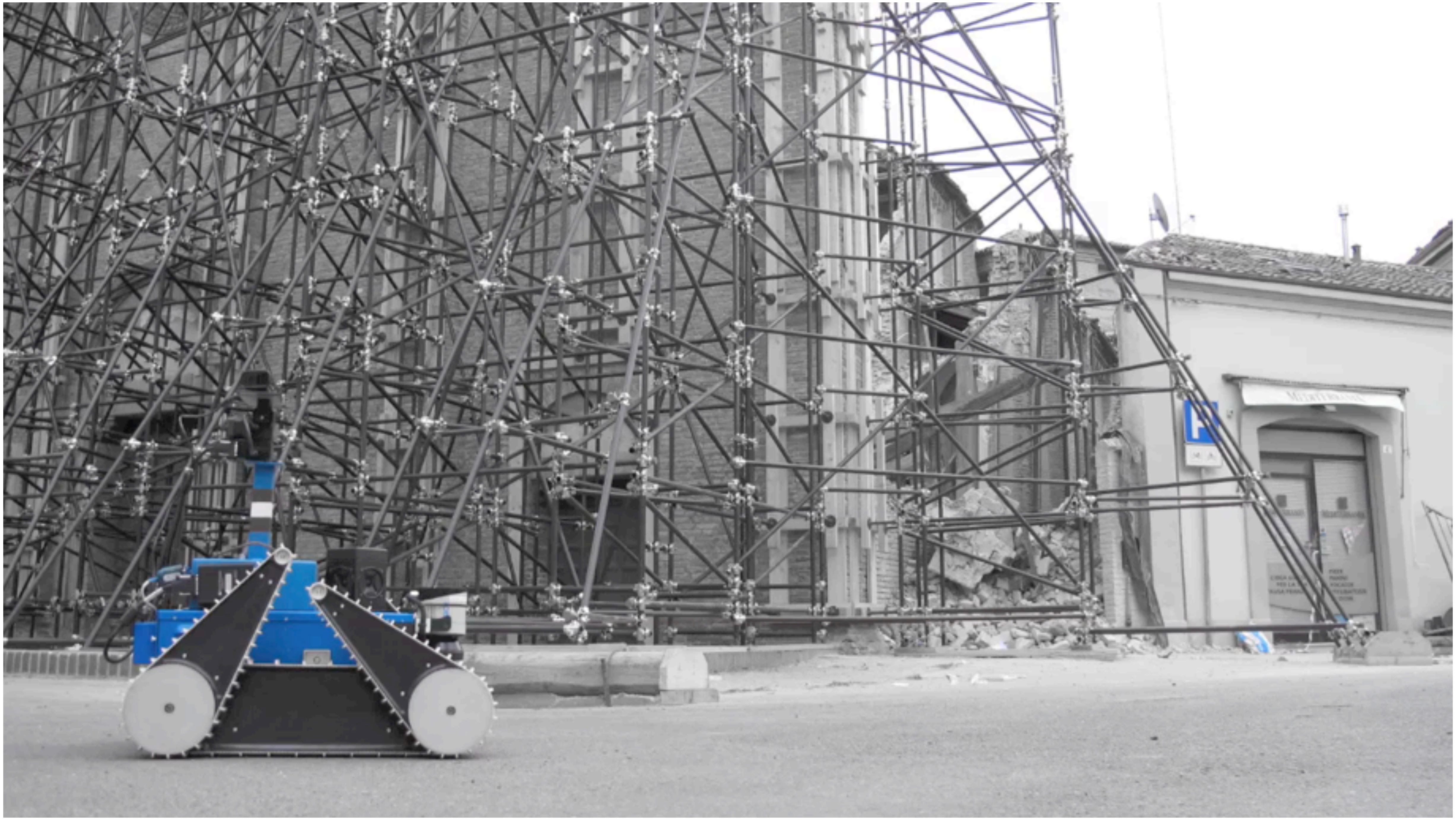
Harsh. Dangerous. **Stressful**.



How does a robot adapt? Act as a team member? help, assist people?

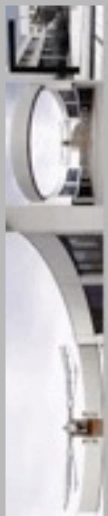




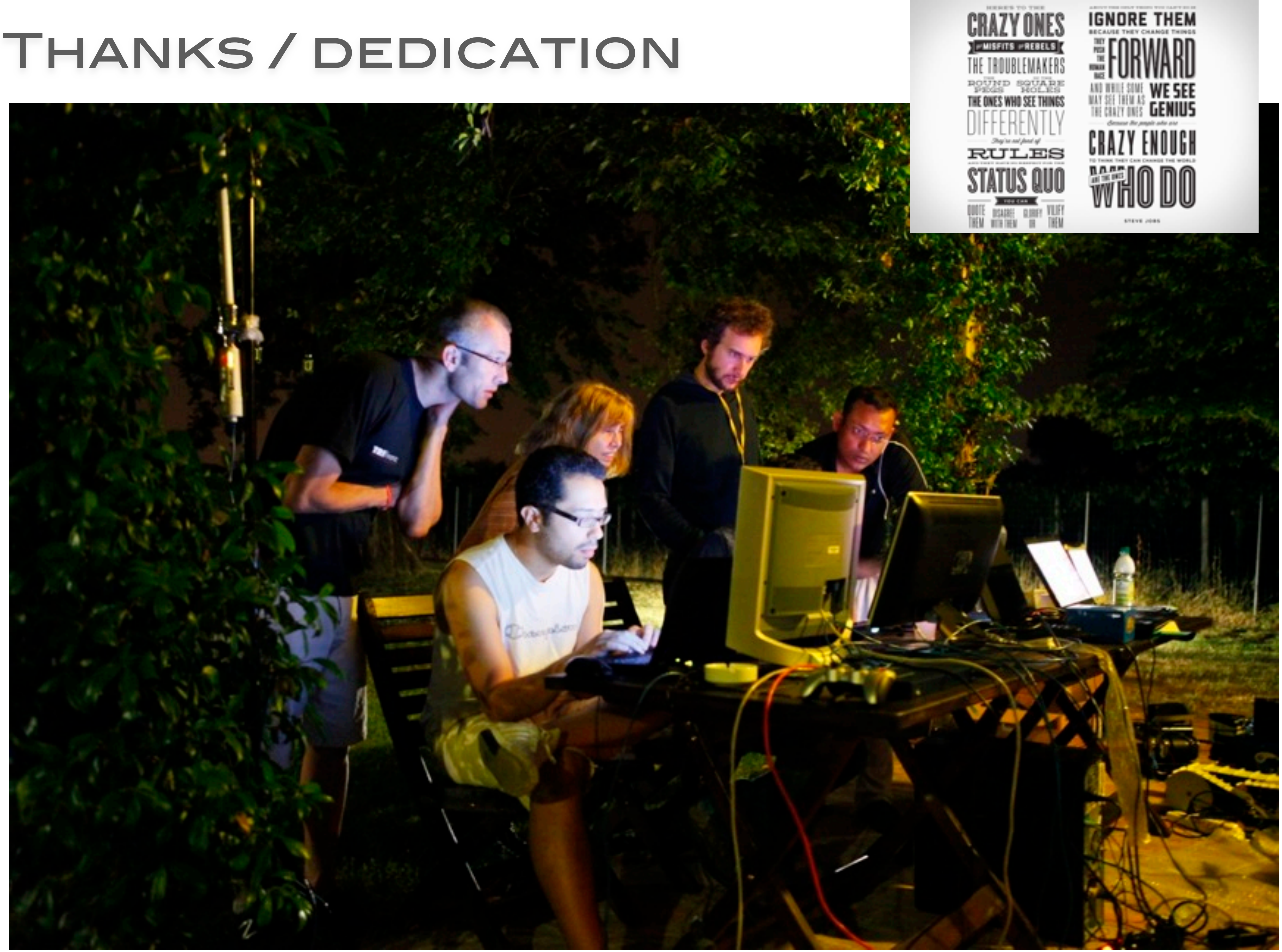


The future

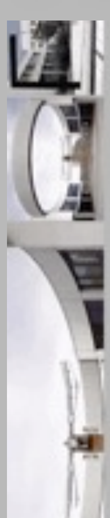
- Adaptation & alignment
  - Adaptation to individual dynamics (cognitive load, situation awareness, stress) and social dynamics (delegation, coordination)
  - Adaptation over long-term interaction to support team growth
  - Alignment between multiple actors in dynamic social contexts
- Real-life teams
  - Develop systems that **really** help people
  - Transparency in team action & interaction
  - Trust between human- and robot team members
  - Growth in HR teams through long-term interaction

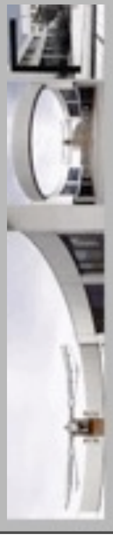


THANKS / DEDICATION



TALKING ROBOTS
Language Technology, DFKI





Thanks / dedication

...

... for making things work, for finishing papers by sunrise, for being jointly eaten alive by mosquitoes while re-installing the system, for wacking away in subterranean “lab” dungeons till we’re done, for coming up with bad ideas and telling me mine are even worse, for doing things nobody thought or even hoped we would be able to do, for always going that little extra 150% “for the team,” for understanding that there is a need for cookies sometimes (or for running), for simply being mad at times, ... “here’s to the crazy ones ... WHO DO.” **Thank you.**

CogSys Cognitive Systems — SEVENTH FRAMEWORK PROGRAMME — COSY — COGX — aliz-e Adaptive Strategies for Sustainable Long-Term Social Interaction — NIFTi Natural Human-Robot Cooperation in Dynamic Environments