DAC-h3: A Proactive Robot Cognitive Architecture to Acquire and Express Knowledge About the World and the Self

Clément Moulin-Frier*, Tobias Fischer*, Maxime Petit, Grégoire Pointeau, Jordi-Ysard Puigbo, Ugo Pattacini, Sock Ching Low, Daniel Camilleri, Phuong Nguyen, Matej Hoffmann, Hyung Jin Chang, Martina Zambelli, Anne-Laure Mealier, Andreas Damianou, Giorgio Metta, Tony J. Prescott, Yiannis Demiris, Peter Ford Dominey, and Paul F. M. J. Verschure

***Abstract*—This paper introduces a cognitive architecture for a humanoid robot to engage in a proactive, mixed-initiative exploration and manipulation of its environment, where the initiative can originate from both the human and the robot. The framework, based on a biologically-grounded theory of the brain and mind, integrates a reactive interaction engine, a number of state-of-the art perceptual and motor learning algorithms, as well as planning abilities and an autobiographical memory. The architecture as a whole drives the robot behavior to solve the symbol grounding problem, acquire language capabilities, execute goal-oriented behavior, and express a verbal narrative of its own experience in the world. We validate our approach in a human-robot interaction experiment with the iCub humanoid robot, showing that the proposed cognitive architecture can be applied in real time within a realistic scenario.**

***Index Terms*—Cognitive Robotics, Distributed Adaptive Control, Human-Robot Interaction, Symbol Grounding, Autobiographical Memory**

I. Introduction

The so-called *Symbol Grounding Problem* (SGP, [1], [2], [3]) refers to how a cognitive agent forms an internal and unified representation of an external word referent from the continuous flow of low-level sensorimotor data generated by its interaction with the environment. In this paper, we focus on solving the SGP in the context of human-robot interaction (HRI), where a humanoid iCub robot [4] acquires and expresses knowledge about the world by interacting with a human partner. Solving the SGP is of particular relevance in HRI, where a repertoire of shared symbolic units forms the basis of an efficient linguistic communication channel between the robot and the human.

To solve the SGP, several questions should be addressed:

- How are unified symbolic representations of external referents acquired from the multimodal information collected by the agent (e.g., visual, tactile, motor)? This is referred to as the *Physical SGP* [**?**], [13].
- How to acquire a shared lexicon grounded in the sensorimotor interactions between two (or more) agents? This is referred to as the *Social SGP* [13], [14].
- How is this lexicon then used for communication and collective goal-oriented behavior? This refers to the functional role of physical and social symbol grounding.

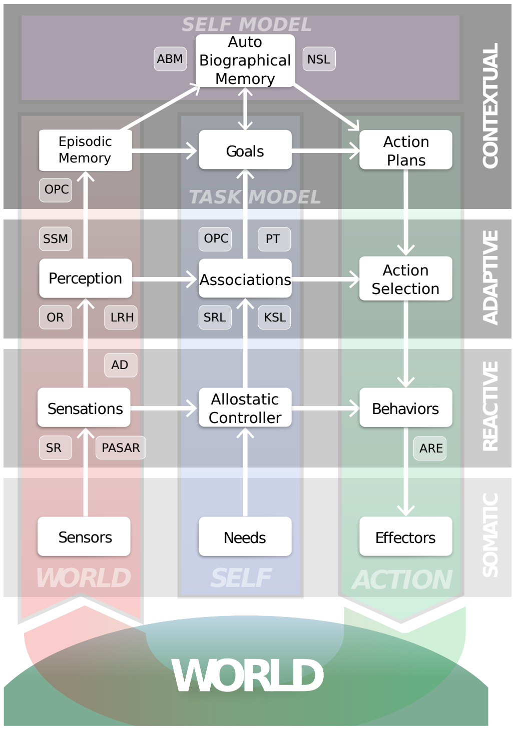

Figure 1. The *DAC-h3* cognitive architecture (see Section III) is an implementation of the DAC theory of the brain and mind (see Section II-B) adapted for HRI applications. The architecture is organized as a layered control structure with tight coupling within and between layers: the somatic, reactive, adaptive and contextual layers. Across these layers, a columnar organization exists that deals with the processing of states of the World or exteroception (left, red), the Self or interoception (middle, blue) and Action (right, green). The role of each layer and their interaction is described in Section III. White boxes connected with arrows correspond to *structural modules* implementing the cognitive modules proposed in the DAC theory. Some of these structural modules rely on *functional modules*, indicated by acronyms in the boxes next to the structural modules. Acronyms refer to the following functional modules. SR: Speech Recognizer; PASAR: Prediction, Anticipation, Sensation, Attention and Response; AD: Agent Detector; ARE: Action Rendering Engine; OR: Object Recognition; LRH: Language Reservoir Handler; SSM: Synthetic Sensory Memory; PT: Perspective Taking; SRL: Sensorimotor Representation Learning; KSL: Kinematic Structure Learning; OPC: Object Property Collector; ABM: Autobiographical Memory; NSL: Narrative Structure Learning.

This paper addresses these questions by proposing a complete cognitive architecture for HRI and demonstrating its abilities on an iCub robot. Our architecture, called *DAC-h3*, builds upon our previous research projects in conceiving

*C. Moulin-Frier and T. Fischer contributed equally to this work.

biologically grounded cognitive architectures for humanoid robots based on the Distributed Adaptive Control theory of mind and brain (DAC, presented in the next section). In [?] we proposed an integrated architecture for generating a socially competent humanoid robot, demonstrating that gaze, eye contact and utilitarian emotions play an essential role in the psychological validity or social salience of HRI (DAC-h1). In [?], we introduced a unified robot architecture, an innovative Synthetic Tutor Assistant (STA) embodied in a humanoid robot whose goal is to interactively guide learners in a science-based learning paradigm through rich multimodal interactions (DAC-h2).

DAC-h3 is based on a developmental bootstrapping process where the robot is endowed with an intrinsic motivation to act and relate to the world in interaction with social peers. This is referred to by Levinson [15] as the *human interaction engine*: a set of capabilities including looking at objects of interest and interaction partners, pointing to these entities [16], curiosity as a desire to acquire knowledge [17], as well as showing, telling and sharing of knowledge with others [16], [18]. This interaction engine drives the robot to proactively control its own acquisition and expression of knowledge, favoring the grounding of acquired symbols by learning multimodal representations of *entities* through interaction with a human partner. In *DAC-h3*, an entity refers to an internal or external referent: it can be either an object, an agent, an action, or a body part. In turn, the acquired multimodal and linguistic representations of entities are recruited in goal-oriented behavior and form the basis of a persistent concept of self through the development of an autobiographical memory and the expression of a verbal narrative.

We validate the proposed architecture following a human-robot interaction scenario where the robot has to learn concepts related to its own body and its vicinity in a proactive manner, as well to express those concepts in goal-oriented behavior. We show a complete implementation running in real-time on the iCub humanoid robot. The interaction depends on the internal dynamics of the architecture, the properties of the environment, as well as the behavior of the human. We analyze a typical interaction in detail and provide videos showing the robustness of our system in various environments (http://clement-moulin-frier.github.io/dac-h3/). Our results show that the architecture autonomously drives the iCub to acquire various concepts about the present entities (objects, humans, and body parts), proactively maintaining the interaction with a human and recruiting those concepts to express more complex goal-oriented behavior.

In Section II we position the current contribution with respect to related works in the field and rely on this analysis to emphasize the specificity of our approach. Our main contribution is described in Section III and consists in the proposition and implementation of an embodied and integrated cognitive architecture for the acquisition of multimodal information about external word referents, as well as a context-dependent lexicon shared with a human partner and used in goal-directed behavior and verbal narrative generation. The experimental validation of our approach on an iCub robot is provided in Section IV, followed by a discussion.

II. Related Works and Principles of the Proposed Architecture

In this section, we review related works on each of the topics related to the features of the proposed cognitive architecture. Those features are: (A) a biologically-grounded cognitive architecture ensuring autonomy, learning and goal-oriented behavior; (B) the ability to solve the SGP by the acquisition of symbols grounded in the physical interaction with the (social) environment; (C) proactivity as a way to improve HRI and self-monitor knowledge acquisition; (D) goal-oriented behavior supported by the learned symbol; as well as (E) autobiographical memory expressed through a verbal narrative. Each of the subsections below presents related works in these topics, and briefly describes how the presented contribution addresses them.

A. Functionally-driven vs. biologically-inspired approaches in social robotics

Conceiving socially interactive robots relies on two main approaches [19]. On the one hand, functionally-designed approaches are based on reverse engineering methods, assuming that a deep understanding of how the mind operates is not a requirement for conceiving socially competent robots (e.g. [20], [21], [22]). On the other hand, biologically-inspired robots are based on theories of natural and social sciences and expect two main advantages of constraining cognitive models by biological knowledge: to conceive robots better understandable by humans since they are built from similar principles, and to provide an efficient experimental benchmark from which the underlying theories can be confronted, tested and refined (e.g. [23], [24]).

The proposed *DAC-h3* cognitive architecture takes advantage of both methods. It is based on an established biologically-grounded cognitive architecture of the brain and the mind (the DAC theory, presented below) that is here adapted for the HRI domain. However, whereas the global structure of the architecture is constrained by biology, the actual implementation of specific modules composing it can sometimes be driven by their functionality, i.e. using state-of-the-art methods from machine learning that are powerful at implementing a particular function without being directly constrained by biological knowledge.

B. Cognitive architectures and the SGP

Another distinction in approaches for conceiving social robots, which is of particular relevance for addressing the SGP, reflects an historical distinction from the more general field of cognitive architectures (or *unified theories of cognition* [25]). Historically, two opposite approaches have been proposed formalizing how cognitive functions arise in an individual agent from the interaction of interconnected information processing modules structured in a cognitive architecture. On one hand, top-down approaches rely on a symbolic representation of a task, which has to be decomposed recursively into simpler ones to be executed by the agent. They mostly rely on methods from symbolic artificial intelligence (from the General

Problem Solver [26] to Soar [27] or ACT-R [28]). Although relatively powerful at solving abstract symbolic problems, top-down architectures are not able to solve the SGP *per se* because they presuppose the existence of symbols and therefore are not suitable for addressing how they are acquired from low-level sensorimotor signals. On the other hand, bottom-up approaches instead implement behavior without relying on complex knowledge representation and reasoning. This is typically the case in behavior-based robotics [29], emphasizing lower-level sensory-motor control loops as a starting point of behavioral complexity as in the Subsumption architecture [30]. Those approaches are not suitable to solve the SGP either because they do not consider symbolic representation as a necessary component of cognition (referred as *intelligence without representation* in [29]). Interestingly, this distinction between top-down representation-based and bottom-up behavior-based approaches still holds in the domain of social robotics [31], [32].

Top-down representation-based approaches rely on modeling psychological aspects of social cognition (e.g. [33]), whereas bottom-up behavior-based approaches emphasize the role of embodiment and reactive control to enable the dynamic coupling of agents [34]. Solving the SGP, both in its physical and social aspects, therefore requires an integration of bottom-up processes for acquiring and grounding symbols in the physical interaction with the (social) environment, with top-down processes for taking advantage of the abstraction, reasoning and communication abilities provided by the acquired symbol system. This has been referred as the *micro-macro loop*, i.e. a bilateral relationship between an emerged symbol system at the macro level and a physical system consisting of communicating and collaborating agents at the micro level [35].

Several contributions in social robotics rely on such hybrid architectures integrating bottom-up and top-down processes (e.g. [38], [37], [36], [**?**]). In [36], an architecture called embodied theory of mind was developed to link high-level cognitive skills to the low-level perceptual abilities of a humanoid and implementing joint attention and intentional state understanding. In [37] or [38], the architecture combines deliberative planning, reactive control, and motivational drives for controlling robots in interaction with humans.

In this paper, we adopt the principles of the *Distributed Adaptive Control* theory of the mind and the brain (DAC, [39], [40]). DAC is a hybrid architecture which posits that cognition is based on the interaction of four interconnected control loops operating at different levels of abstraction (see Figure 1). The first level is called the *somatic layer* and corresponds to the embodiment of the agent within its environment, with its sensors and actuators as well as the physiological needs (e.g. for exploration or safety). Extending bottom-up approaches with drive reduction mechanisms, complex behavior is bootstrapped in DAC from the self-regulation of an agent's physiological needs when combined with reactive behaviors (the *reactive layer*). This reactive interaction with the environment drives the dynamics of the whole architecture [41], bootstrapping learning processes for solving the physical SGP (the *adaptive layer*) and the acquisition of higher-level cognitive representations such as abstract goal selection, memory and planning (the *contextual layer*). These high-level representations in turn modulate the activity at the lower levels via top-down pathways shaped by behavioral feedback. The control flow in DAC is therefore distributed, both from bottom-up and top-down interactions between layers, as well as from lateral information processing into the subsequent layers.

C. Solving the SGP

Other contributions have focused on the grounding of a lexicon from the physical interaction of a robot with its environment. Since the pioneering paradigm of "language games" proposed in [42], a number of multi-agent models have been proposed showing how particular properties of language can self-organize out of repeated dyadic interactions between agents of a population (e.g. [43], [44]). In the domain of HRI, contributions have focused on lexicon acquisition through the transfer of sensorimotor and linguistic information from the interaction between a teacher and a learner through imitation [45], action [46], [47] or active exploration [48]. In all these contributions, solving the SGP requires integrating multimodal information about external world entities (*physical SGP*) with linguistic labels acquired or negotiated through the interaction with social peers (*social SGP*).

In the proposed *DAC-h3* architecture, entities refer to either object, agents, actions or body parts. Visual, tactile, motor and linguistic information about the present entities is collected proactively through reactive control loops triggering knowledge acquisition and expression behaviors in interaction with a human partner. An entity is therefore represented internally in the robot's memory as the association between multimodal representations and linguistic labels.

D. Autonomous exploration and proactive behavior in HRI

Significant progress has been made in allowing robots to interact with humans, for example in learning shared plans [5], [6], [7], learning to imitate actions [8], [9], [10], and learning motor skills [11] that could be used for engaging in joint activities. In most of these interactions, the human is in charge and the robot is following the human's lead: the choice of the concept to learn is left to the human and the robot must identify it. In this case, the robot must solve the referential indeterminacy problem described by Quine [12], where the robot language learner has to extract the external concept that was referred to by a human speaker. However, acquiring symbols by interacting with other agents is not only an unidirectional process of information transfer between a teacher and learner [15].

Autonomous exploration and proactive behavior allow robots to take the initiative in exploring their environment [49] and interacting with people [50]. The benefit of these abilities for knowledge acquisition has been demonstrated in several HRI experiments. In [48], it is shown how a combination of social guidance and intrinsic motivation improve the learning of object visual categories in HRI. In [51], planning conflicts due to the uncertainty of the detected human's intention are resolved by proactive execution of the corresponding task that optimally reduces the system's uncertainty. In [52], the task is to acquire human-understandable labels for novel objects and learning how to manipulate them. This is realized through a mixed-initiative interaction scenario and it is shown that proactivity improves the predictability and success of human-robot interaction.

A central aspect of the *DAC-h3* architecture is the robot's ability to act proactively in a mixed-initiative scenario. This allows the self-monitoring of the robot's own knowledge acquisition process, instead of only relying on the human's initiative. Interestingly, proactivity in a sense reverses the referential indeterminacy problem mentioned above, by shifting the responsibility of solving ambiguities to the agent who is endowed with the adequate prior knowledge to solve it, i.e., the human in a HRI context. The robot is now in charge of the concepts it wants to learn, and can use joint attention behaviors to guide the human toward the knowledge it wants to acquire. In the proposed system, this is realized through a set of behavioral control loops, self-regulating knowledge acquisition, and proactively requesting missing information about entities from the human partner.

E. Language learning, autobiographical memory and narrative expression

Several works address the ability of language learning in robotics. The cognitive architecture of iTalk [53] focuses on modeling the emergence of language by learning about the robot's embodiment, learning from others, as well as learning linguistic capability. Cangelosi *et al.* [54] propose that action, interaction and language should be considered together as they develop in parallel, and one influences the others. Antunes *et al.* [55] assume that language is already learned, and address the issue that linguistic input typically does not have a one-to-one mapping to actions. They propose to perform reasoning and planning on three different layers (low-level robot perception and action execution, mid-level goal formulation and plan execution, and high-level semantic memory) to interpret the human instructions. Similarly, [56] proposes a system to recognize novel objects using language capabilities in one shot. In these works, language is typically used to understand the human and perform actions, but not necessarily to talk about past events which the robot has experienced.

Various works investigate the expression of past events by developing narratives based on acquired autobiographical memories [57], [58], [59]. In [59], a user study is presented which suggests that a robot's narrative allows humans to get an insight to long term human-robot interaction from the robot's perspective. The method in [58] takes user preferences into account when referring to past interactions. Similarly to our framework, it is based on the implementation and cooperation between both episodic and semantic memories with a dialog system. However, no learning capabilities (neither language nor knowledge) are introduced by the authors.

In the proposed *DAC-h3* architecture, the acquired lexicon allows the robot to execute action plans for achieving goal-oriented behavior from human speech requests. Relevant information throughout the interaction of the robot with humans is continuously stored in an autobiographical memory used for the generation of a narrative self, i.e., a verbal description of the own robot's history over the long term (able to store and verbally describe interactions from a long time ago, e.g. several months ago).

In the next section, we describe how the above features are implemented in a coherent cognitive architecture relying on functional modules which are implemented as YARP [60] modules running in real-time on the iCub robot.

III. The DAC-h3 Cognitive Architecture

This section presents the *DAC-h3* architecture in detail, which is an instantiation of the DAC architecture for human-robot interaction. The proposed architecture provides a general framework for designing autonomous robots which act proactively for 1) maintaining social interaction with humans, 2) bootstrapping the association of multimodal knowledge with its environment that further enrich the interaction through goal-oriented action plans, and 3) express a verbal narrative. It allows a principled organization of various *functional modules* into a biologically grounded cognitive architecture.

In *DAC-h3*, the *somatic layer* consists of an iCub humanoid robot equipped with advanced motor and sensory abilities for interacting with humans and objects. The *reactive layer* ensures the autonomy of the robot through drive reduction mechanisms implementing proactive behaviors for acquiring and expressing knowledge about the current scene. The *adaptive layer* learns high-level multimodal representations (visual, tactile, motor and linguistic) for the categorization of entities (objects, agents, actions and body parts) and associates them in unified representations. Finally, the *contextual layer* deals with goal representation and action planning, as well as the formation of an autobiographical memory of the robot that can be expressed in the form of a verbal narrative.

The complete *DAC-h3* architecture is shown on Figure 1. It is composed of *structural modules* reflecting the cognitive modules proposed by the DAC theory. Each *structural module* possibly relies on one or more *functional modules* implementing more specific functionalities (e.g. dealing with motor control, object perception, and scene representation). In this section, we describe each *structural module* layer by layer, as well as their interaction with the *functional modules*.

A. Somatic layer

The *somatic layer* corresponds to the physical embodiment of the system. We use the iCub robot, an open source humanoid platform developed for research in cognitive robotics [4]. The iCub is a 104 cm tall humanoid robot with 53 degrees of freedom (DOF). It has two dexterous hands with 19 under-actuated joints and 9 DOF each. The robot is equipped with cameras in its articulated eyes allowing stereo vision, and tactile sensors in the fingertips, palms of the hand, arms and torso. The iCub is augmented with an external RGB-D camera above the robot head for agent detection and skeleton tracking. Finally, an external microphone and speakers are used for speech recognition and synthesis, respectively.

The *somatic layer* also contains the physiological *needs* of the robot that will drive its reactive behaviors, as described in the following section on the *reactive layer*.

B. Reactive layer

Following DAC principles, the *reactive layer* oversees the self-regulation of the internal *drives* of a cognitive agent from the interaction of sensorimotor control loops. The drives aim at self-regulating internal state variables (the *needs* of the *somatic layer*) within their respective homeostatic ranges. In biological terms, such an internal state variable could for example reflect the current glucose level in an organism, with the associated homeostatic range defining the minimum and maximum values of that level. A drive for eating would then correspond to a self-regulation mechanism where the agent actively searches for food whenever its glucose level is below the homeostatic minimum, and stops eating even if food is present whenever it is above the homeostatic maximum. A drive is therefore defined as the real-time control loop triggering appropriate behaviors whenever the associated internal state variable goes out of its homeostatic range, as a way to self-regulate its value in a dynamic and autonomous way.

In the social robotics context that is considered in this paper, the drives of the robot do not reflect biological needs as above, but are rather related to knowledge acquisition and expression in social interaction. At the foundation of this developmental bootstrapping process is the intrinsic motivation to interact and communicate. As described by Levinson [15], a part of the human interaction engine is a set of capabilities that include the motivation to interact and communicate through universal (language independent) manners; including looking at objects of interest and at the interaction partner, and pointing to these objects. These reactive capabilities are built into the *reactive layer* of the architecture forming the core of the *DAC-h3* interaction engine. These interaction primitives allow the *DAC-h3* system and the human to share attention around specific entities (body parts, objects, or agents), bootstrapping learning mechanisms in the *adaptive layer* that associate visual, tactile, motor and linguistic representations of entities (see next section on the *adaptive layer*).

Currently the architecture implements the following two drives, however *DAC-h3* is designed in a way that facilitates the addition of new drives for further advancements. First, a *drive for knowledge acquisition* provides the iCub with an intrinsic motivation to acquire new knowledge about the current scene. The internal variable associated with this drive is modulated by the number of entities (object, agents, body parts, etc.) in the current scene with missing information (e.g. unknown name, or missing property). The self-regulation of this drive is realized by proactively requesting the human to provide missing information about entities, for instance their name via speech, synchronized with gaze and pointing; or asking the human to touch its skin associated to a specific body part. Second, a *drive for knowledge expression* allows the iCub to proactively express its acquired knowledge by interacting with the human and objects. The internal variable associated with this drive is modulated by the number of entities in the current scene without missing information. The self-regulation is then realized by triggering actions toward the known entities, synchronized with verbal descriptions of those actions (e.g. pointing towards an object or moving a specific body part, while referring verbally to the considered entity).

The implementation of these drives is realized through the three structural modules described below, interacting with each other as well as with the surrounding layers: *1) sensations*, *2) allostatic controller*, and *3) behaviors* (see Figure 1).

1) Sensations: The *sensations* module pre-processes the data acquired by the sensors at the *somatic level*: the positions of entities present in the current scene (objects and the human), their visual salience, skeleton information from the RGB-D camera, as well as information indicating the presence or absence of the entity properties (linguistic label, or in case of body parts touch and motor information). The *sensations* module relies on several functional modules described below. *Agent detector* and the *prediction, anticipation, sensation, attention and response* module maintain information relative to the position of each entity and its respective salience, as well as *speech recognition* which deals with speech-to-text processing to provide relevant information for meaning extraction in the *adaptive layer*.

The *agent detector* is a functional module processing information gathered by the RGB-D camera and concerning a human agent present in the scene. This includes the location of the body parts of the human (of interest in this paper is the head location for gazing at the partner, the location of the hands for detecting pointing actions of the human, as well as the face image of the agent for face recognition.

The *prediction, anticipation, sensation, attention and response* (PASAR, [**?**]) functional module analyses the salience of entities in the world. The salience of an object is increased through acceleration of the entity, its appearance and disappearance, the human pointing to the object; and decreases when the iCub looks at the entity. The module receives information about properties of detected entities (position, color, presence), and about the human interaction with them (pointing to, waving to, etc.). After performing analysis of the salience of the object, the system generates the updated salience for the entities. This can be used to guide the gaze to the most pertinent objects.

The *speech recognition* functional module extracts text from human speech sensed by a microphone, using the Microsoft$^{\text{TM}}$ speech API. With a given grammar provided by different modules or given by default, this module returns the sentence heard by the robot in textual form. This will be further processed in the *adaptive layer* described in the next section to extract meaning from those sentences.

2) Allostatic Controller: In many situations, several drives which may conflict with each other, can be activated at the same time (in the case of this paper, the drive for knowledge acquisition and the drive for knowledge exploration). Such possible conflicts can be solved through the concept of an *allostatic controller* [61], [62], defined as a set of simple homeostatic control loops and dealing with their scheduling to ensure an efficient global regulation of the internal state variables. According to the internal state of the robot and the output of the *sensations* module, the *allostatic controller* updates the drive levels in real-time and has the role of deciding which drive to regulate at the current time. For example, the knowledge acquisition drive is modulated by the amount of unknown information about the entities present in the current scene, whereas the knowledge expression drive is modulated by the amount of already acquired information about the entities present in the current scene. The decision of which drive to follow depends on several factors: the distance of each drive level to their homeostatic boundaries, as well as predefined drive priorities (in *DAC-h3*, knowledge acquisition has priority over knowledge expression, which results in a curious personality).

3) Behaviors: To regulate the aforementioned drives, the *allostatic controller* is connected to the *behaviors* module, and each drive is linked to corresponding behaviors which are supposed to bring it back into its homeostatic range whenever needed. The drive for knowledge acquisition is regulated by requiring information about entities through coordinated behaviors. Those behaviors depend on the type of the considered entity:

- In case of an object, the robot produces speech (e.g. "What is this object?") while pointing and gazing at the unknown object.
- In case of an agent, the robot produces speech (e.g. "Who are you?") while looking at the unknown human.
- In case of a body part, the robot either asks for the name (e.g. "How do you call this part of my body?") while moving it or, if the name is already known from a previous interaction, asks the human to touch the body part while moving it (e.g., "Can you touch my index while I move it, please?").

The multimodal information collected through these behaviors will be used to form unified representation of entities in the *adaptive layer* (see next section).

The drive for knowledge expression is regulated by executing actions towards known entities, synchronized with speech sentences parameterized by the entities' linguistic labels acquired in the *adaptive layer* (see next section).

Motor actions are realized through the *action rendering engine* (ARE [63]) functional module, executing complex actions (e.g. push, reach, take, look) in terms of primitive motor movements in both the joint and the Cartesian space. The ARE system takes as input the label of the requested action (e.g. push) along with the 3D coordinates of the object (provided by the *sensations* module) over which the action needs to be executed. The whole trajectory gets decomposed in multiple trajectories via points whose intermediate movements are resolved by means of a non-linear constraints optimization, and then performed by a multi-referential operational controller as described in [63]. The robot executes the motor actions moving its limbs in a coordinated human-like fashion.

Language production abilities are also present in the reactive layer, implemented in the form of predefined grammars (for example naming an object or requesting information about it). Semantic words associated to entities are not present at the reactive level, but are provided from the learned association operating in the *adaptive layer* (see next section). The *iSpeak* module implements a bridge between the iCub and a voice synthesizer (e.g. Festival[1], Acapela[2]) by synchronizing the produced utterance from a string with the LEDs of the iCub's mouth, producing lips movements to realize a more vivid interaction [64].

The positive influence of such a drive regulation mechanism on the acceptance of the HRI by naive users has been demonstrated in previous papers [65], [66]. In this paper, the focus is on its integration with an *adaptive* and *contextual layer*, as described in the next two sections, providing a complete cognitive architecture for HRI.

C. Adaptive layer

The *adaptive layer* oversees the acquisition of a state space of the agent-environment interaction by binding visual, tactile, motor and linguistic representations of entities. It integrates functional modules for maintaining an internal representation of the current scene, visually categorizing entities, recognizing and sensing body parts, extracting linguistic labels from human speech, and learning associations between multimodal representations. They are grouped in three structural modules described below: *perceptions*, *associations* and *action selection* (see Figure 1).

1) Perceptions: The *object recognition* functional module [67] is used to learn the categorization of objects directly from the visual information given by the iCub eyes. It provides a complete chain to address the real-world object recognition problem by coding the images with resort to the most recent deep convolutional networks, and successively applying a support vector machine linear classification. The input to this system are images acquired from the robot cameras containing the objects we aim to recognize. Objects are segmented out of the background relying on the local binary pattern technique; their bounding boxes are then fed to the learning module for the recognition stage. The whole chain can run in near real-time. The output of the system consists of the 2D (in the image plane) and 3D (in the world frame) positions of the identified objects along with the corresponding classification scores as stored in the *objects properties collector* memory (explained below).

The transformation of human speech into meaning (comprehension of narrative discourse) and vice versa (production of narrative discourse) is realized by the *language reservoir handler* (LRH). These models (comprehension and production of narrative discourse) are recurrent neuronal networks based on reservoir computing [68], [69], [70]. The model of narrative discourse production (meaning to speech) receives a representation of meaning and generates the corresponding sentence. The narrative discourse comprehension model is the opposite of the production model. It receives a sentence and produces the representation of the corresponding meaning, which is based on the theory of Goldberg [71] who proposes a tight correspondence between the structure of perceptual events that are basic to human experience, and constructions for the corresponding basic sentence types. In a simplistic way, this cor-

[1] http://www.cstr.ed.ac.uk/projects/festival/

[2] http://www.acapela-group.com/

responds to saying that the meaning of a sentence is contained in the semantic words that compose it. Thereby, we represent the meaning in term of PAOR: predicate(arguments) where arguments correspond to thematic roles (agent,object,recipient). These thematic roles called "Open Class Words" (OCW) are at the interface of syntax and semantics. They are a first level of "poor" semantics. From a syntactic point of view, thematic roles are linked to grammatical functions (such as *to, the, in, on, a*) also called "Closed Class Words" (CCW). The first role we can define is the "agent" role (such as *I, John, iCub*). It typically characterizes a person or entity with intelligence and cognitive properties that initiates or undergoes an action described by a verb (such as *put, push, feel*). We then find the notion of "patient" or "recipient" (*iCub gives the box to John*, in this example, the recipient is *John*), which characterizes the entity that undergoes the action described by the predicate. The "predicate" thematic role represents the key thematic role of an action. In case of narrative structure, we can also find narrative function words like "before", "then", etc. This meaning representation in term of PAOR corresponds to the format of data contained in the *autobiographical memory* described in the next section (*contextual layer*).

The *synthetic sensory memory* (SSM) module provides abilities for pattern learning, recall, pattern completion and association, as defined by experimental psychologists and neuroscientists [86]. The functionality in SSM draws inspiration from the role of the hippocampus by fusing multiple sensory input streams and representing them in a latent feature space, which emerges as *Deep Gaussian Processes* [87] are used as the underlying technology. SSM takes various data as input as provided from the *sensations* module, e.g. vision, auditory, and touch information. SSM is capable of voluntary recall, involuntary recall, as well as imagination. During voluntary recall, SSM performs classification of incoming sensory data and returns a label along with an uncertainty measure corresponding to the returned label. SSM is also capable of imagining novel inputs or reconstructing previously encountered inputs and sending the corresponding generated sensory data which allows for the replay of memories using the visual memory inspector [88]. In *DAC-h3*, SSM is currently employed for face recognition and action recognition using a fusion of RGB-D data and object location data. In terms of action recognition, SSM has been trained to automatically segment and recognize the following actions: push, pull, lift, drop, wave, and point, while also actively recognizing if the current action is known or unknown.

2) Associations: The *associations* structural module produces unified representations of entities by associating the multimodal categories formed in the *perception* module. Those unified representations are formed in the *objects properties collector* (OPC), a functional module storing all information associated with a particular entity at the present moment in a proto-language format as detailed in [65]. An entity is defined as a concept which can be manipulated, and is thus the basis for emerging knowledge. In *DAC-h3*, each entity has a name associated, which might be unknown if the entity has been discovered but not yet explored. More specifically, higher level entities such as objects, body parts and agents have additional intrinsic properties. For example, an object also has a location and dimensions associated to it. Furthermore, whether the object is currently present is encoded as well, and if so, its saliency value (as computed by the PASAR module described in Section III-B). On the other hand, a body part is an entity which contains a proprioceptive property (i.e. a specific iCub's joint), and a tactile information property (i.e. which tactile sensor are associated to it). Thus, the OPC allows integrating multiple modalities of one and the same entity to ground the knowledge about the self, other agents, and objects, as well as their relations. Relations can be used to link several instances in an ontological model (see III-D1 *Episodic Memory*).

Learning the multimodal associations that form the internal representations of entities relies on the behavior generated by the knowledge acquisition drive operating at the *reactive level* (see previous section). Multimodal information about entities generated by those behaviors is bound together by registering the acquired information in the specific data format used by the OPC. For instance, the *language reservoir handler* module described above deals with speech analysis to extract entity labels from human replies (e.g. "this is a cube"; {P:is, A:this, O:cube, R:$\emptyset$}). The extracted labels are associated with the acquired multimodal information which depends on the entity type: visual representations generated by the *object recognition* module in case of an object or *agent detector* in case of an agent, as well as motor and touch information in case of a body part.

The associations of representations can also be applied to the developmental robot itself (instead of external entities as above), to acquire motor capabilities or to learn the links between motor joints and skin sensors of its body [72]. Learning self-related representations of the robot's own body schema is realized by the *sensorimotor representation learning* functional module dedicated to forward model learning by means of an online heterogeneous ensemble of predictors [73]. This learning method achieves predictions which are more accurate compared with single models' alternatives by combining multiple predictors of different types. The ensemble includes echo-state networks [74], online echo state Gaussian processes [75] and locally weighted projection regression models [76]. The system receives sensory data collected from the robot's sensors (e.g. cameras, skin, joint encoders). Sensory predictions are then produced using the self-learned forward model implemented by the proposed ensemble method.

The *kinematic structure learning* functional module [77], [78] estimates an articulated kinematic structure of arbitrary objects (including the robot's body parts, and humans) using local feature point trajectories extracted from the visual input videos of the iCub eye cameras. By combining estimated motion and skeleton information, it generates elaborate and plausible kinematic structures. Based on the estimated articulated kinematic structures [77], we also allow the iCub to anchor two objects' kinematic structure joints by observing their movements [78]. For this goal, we formulated the problem of finding corresponding kinematic joint matches between two articulated kinematic structures via hypergraph matching, whilst being accurate and plausible under appearance and motion variations. The similarity measures consider structural topology (first order), kinematic correlation (second order) and combinatorial motion (third order) similarities simultaneously. This allows the iCub to infer correspondences between its own body parts (its left arm and its right arm), as well as between its own body and the body of the human as retrieved by the *agent detector*.

Finally, the *perspective taking* functional module [79] enables the robot to reason about the state of the world from the partner's perspective. This is important in situations where the views of the robot and the human diverge, for example due to objects which are hidden to the human but visible to the robot. Interestingly, it was proposed that there are two separate processes involved in perspective taking depending on the difficulty of the task [80]. We follow this proposal and solve rather simple tasks such as determining whether an object is visible to the human using line of sight tracing. Complex tasks such as visually imagining the world from the human's point of view are solved by a mental rotation. For this, the point cloud acquired from the RGB-D camera is rotated such that the origin coincides with the head of the human. Furthermore, the mental rotation allows algorithms to reason as if the input data was acquired from an egocentric perspective, which can be used for left/right judgments of object locations. As input, the system receives the object locations; the estimated head pose of the partner; and the point cloud acquired from the RGB-D camera. The system produces an output that characterizes the visibility of each object, the spatial location of each object (left / right of human), as well as a reconstructed view from the human's perspective. This can then be used when reasoning about the object which is referred to by the human.

3) Action Selection: The *action selection* module uses the information from *associations* to provide context to the *behaviors* module at the reactive level. This context corresponds to entity names which are provided as parameters to the *behaviors* module, for instance pointing at a specific object or using the object linguistic label in the parameterized grammars defined at the reactive level. This module also deals with the scheduling of action plans from the *contextual layer* according to the current state of the system as explained in the next section.

D. Contextual layer

The *contextual layer* deals with higher-level cognitive functions that extend the time horizon of the cognitive agent, such as an episodic memory, goal representation, planning and the formation of a persistent autobiographical memory of the robot interaction with the environment. These functions rely on the unified representations of entities acquired at the *adaptive level*. The *contextual layer* consists of three functional modules that are described below: *1) episodic memory*, *2) goals and action plans*, and *3) autobiographical memory* used to generate a narrative structure.

1) Episodic Memory: The *episodic memory* relies on advanced functions of the *object property collector* (OPC) to store and associate information about entities in a uniform format based on the interrogative words "who", "what", "where", "when", "why" and "how" called an H5W data structure [65]. It is used for goal representation and as elements of the *autobiographical memory*. Formalizing the content and evolution of a scene requires the combination of perceptual, symbolic and rule based reasoning in a single unified framework. Such processes will generate information about who is acting, what they are doing, where and when it happens, and this will give cues about why it is happening. H5W have been argued to be the main questions any conscious being must answer in order to survive in the world [81], [82].

The concept of *relations* is the core of the H5W framework. It links up to five concepts and assigns them with semantic roles to form a solution to the H5W problem. We define a *relation* as a set of five edges connecting those nodes in a directed and labeled manner. The labels of those edges are chosen so that the *relation* models a typical sentence from the English grammar of the form: Relation → Subject Verb [Object] [Place] [Time]. The brackets indicate that the complements are facultative; the minimal relation is therefore composed of two entities representing a subject and a verb.

2) Goals and action plans: Goals can be provided to the iCub from human speech, and a meaning is extracted by the *language reservoir handler*, forming the representation of a goal in the *goals* module. Each goal consequently refers to the appropriately predefined action plan, defined as a state transition graph with states represented by nodes and actions represented by edges of the graph. The *action plans* module extracts sequences of actions from this graph, with each action being associated with a pre- and a post-condition state. Goals and action plans can be parameterized by the name of a considered entity. For example, if the human asks the iCub to take the cube, this loads an action plan for the goal "Take an object" which consists in two actions: "Ask the human to bring the object closer" and "Pull the object". In this case, each action is associated with a pre- and post-condition state in the form of a region in the space where the object is located. In the *action selection* module of the adaptive layer, the plan is instantiated toward a specific object according to the knowledge retrieved from the *associations* module (e.g. allowing to retrieve the current position of the cube). The minimal sequence of actions achieving the goal is then executed according to the perceived current state updated in real-time, repeating each action until its post-condition is met (or giving up after a predefined timeout).

Although quite rigid in its current implementation, in the sense that action plans are predefined instead of being learned from the interaction, this planning ability allows closing the loop of the whole architecture, where drive regulation mechanisms at the *reactive layer* can now be bypassed through contextual goal-oriented behavior.

3) Autobiographical Memory: The *autobiographical memory* (ABM [83], [84], [85]) collects long term information (days, months, years) about interactions. It stores data (e.g. objects locations, human presence) from the beginning (i.e. to extract pre-conditions) to the end (i.e. to extract the effect) of an episode by taking snapshots of the environmental information from the *episodic memory*. This allows the generation of high level concepts extracted by knowledge-based reasoning. The memory information itself is stored in a SQL database.

The ABM is bio-motivated based on the human declarative long term memory situated in the medial temporal lobe, and the distinction between facts and events [83], [84]. In addition, the ABM captures continuous information during an episode (e.g. images from the camera, joints values). This can be used by reasoning modules which focus on the action itself, leading to the production of a procedural memory (e.g. through learning from motor babbling or imitation) [85].

The *narrative structure learning module* builds on the language processing and ABM capabilities. Narrative structure learning occurs in three phases: First the iCub acquires experience in a given scenario. This generates the meaning representation in the ABM. The iCub then formats each story using an IGARF graph (for Initial Goal Action Result Final) according to the different states, actions and results of actions encountered [70]. The human then provides a narration (that is understood using the reservoir system explained in Section III-C) for the scenario. By mapping the events of the narration to the event of the story, the robot can extract the meaning of different discourse functions words (such as because). It can thus automatically generate the corresponding form-meaning mapping that defines the individual grammatical constructions, and their sequencing that defines the narrative construction for a new narrative.

E. Layer Synergy

The complete system described in this section therefore integrates several state-of-the-art algorithms for cognitive robotics, and integrates them into a structured cognitive architecture grounded in the principles of the DAC theory. This allows a complex control of the iCub robot proactively interacting with humans trough drive reduction mechanisms in the *reactive layer*, which allows the bootstrapping of adaptive learning of multimodal representations about entities in the *adaptive layer*. Those representations form the basis of an episodic memory for goal-oriented behavior through planning in the *contextual layer*. The life-long interaction of the robot with humans continuously feed an autobiographical memory able to retrieve past experience from request and to express it verbally in a narrative.

IV. Experimental Results

This section validates the cognitive architecture described in the previous section on a real demonstration with an iCub humanoid robot interacting with objects and a human. We first describe the experimental setup, then the autonomous and goal-oriented abilities provided to the robot. Finally we analyze the scenario progression in detail.

The code for reproducing this experiment on any iCub robot is available open-source on GitHub at https://github.com/robotology/wysiwyd. It consists of all modules described in the last section implemented in either C++ or Python, and relies on the YARP middleware [60] for defining their connections and ensuring their parallel execution in real-time.

A. Experimental setup

We consider an HRI scenario where the iCub and a human face each other with a table in the middle and objects placed on it. The surface of the table is divided in three distinct areas, as shown in Figure 2:

1) an area which is only reachable by the iCub (I),
2) an area which is only reachable by the human (H), and
3) an area which is reachable by both agents (S for *Shared*).

The behaviors available to the iCub are the following:

- “Acquire missing information about an entity”, which is described in more detail in Section IV-B1.
- “Express the acquired knowledge”, which is described in more detail in Section IV-B2.
- “Move an object on the table”, either by pushing it from the region I to S or pulling it from the region S to I,
- “Ask the human to move an object”, either asking to push the object from region H to S or asking to pull it from region S to H.
- “Show learned representations on screen” while explaining what is being shown, e.g. displaying the robot kinematic structure learned from a previous arm babbling phase.
- “Interact verbally with the human” while looking at her/him. This is used for replying to some human requests as described in Section IV-C below.

These behaviors are implemented in the *behaviors* module and can be triggered from two distinct pathways as shown in Figure 1. The behaviors for acquiring and expressing knowledge are triggered through the drive reduction mechanisms implemented in the *Allostatic Controller* (Section III-B). The remaining behaviors are triggered from the *Action Selection* module (Section III-C), scheduling action sequences from the *Goals* and *Action Plans* modules (Section III-D. We describe these two pathways in the two following subsections.

B. Autonomous behavior

Two drives for knowledge acquisition and knowledge expression implement the interaction engine of the robot (see Section III-B). They regulate the knowledge acquisition process of the iCub and proactively maintain the interaction with the human. The generated sensorimotor data feeds the adaptive layer of the cognitive architecture to acquire multimodal information about the present entities (see Section III-C). In the current experiment, the entities are objects on the table, body parts (fingers of the iCub), interacting humans, and actions. The acquired multimodal information depends on the considered entity. Object representations are based on visual categorization and stereo-vision based 3D localization performed by the *object recognition* functional module. Body part representations associate motor and touch events. Agents and actions representations are learned from visual input in the *synthetic sensory memory* module presented in Section III-C1. Each entity is also associated with a linguistic label learned by self-regulating the two drives detailed below.

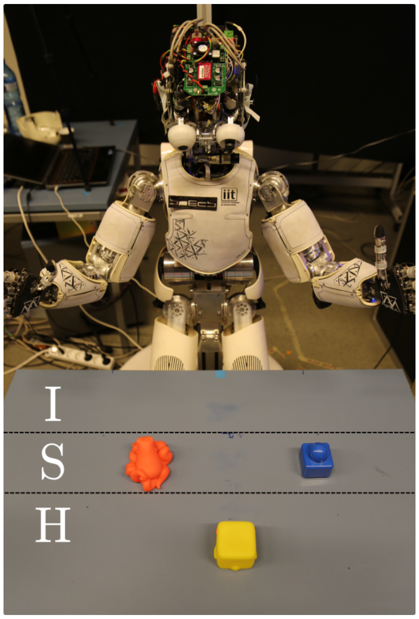

Figure 2. The setup consists of an iCub robot interacting with objects on the table and a human in front of it. The three regions of the table mentioned in the text (I, S and H) are indicated (separated by horizontal lines).

1) Drive to acquire knowledge: This drive maintains a curiosity-driven exploration of the environment by proactively requesting to the human information about the present entities, e.g. the name of an object or the touch of a body part. The drive level decays proportionally to the amount of missing information about the present entities (e.g. the unknown name of an entity). When below a given threshold, it triggers a behavior following a generic pattern of interaction, instantiated according to the nature of the knowledge to be acquired. It begins with a behavior to obtain a joint attention between the human and the robot toward the entity that the robot wants to learn about. After the attention has been attracted toward the desired entity, the iCub asks for the missing information (e.g. the name of an object or of the human, or in case of a body part the name and touch information) and the human replies accordingly. In a third step, this information is passed to the adaptive layer and the knowledge of the robot is updated in consequence.

Each time the drive level reaches the threshold, an entity is chosen in a pseudo-random way within the set of perceived entities with missing information, with a priority to request the name of a detected unknown human partner. Once a new agent enters the scene, the iCub asks for her/his name, which is stored alongside representations of its face in the *synthetic sensory memory* module. Similarly, the robot stores all objects it has previously encountered in its *episodic memory* implemented by the *object property collector* module. When the chosen entity is an object, the robot asks the human to provide the name of interest while pointing at it. Then, the visual representation of the object computed by the *object recognition* module is mapped to the name. When the chosen entity is a body part (left hand fingers), the iCub first raises its hand and moves a random finger to attract the attention of the human. Then it asks for the name of that body part. This provides a mapping between the robot's joint identifier and the joint's name. This mapping can be later extended to also include tactile information, by asking the human to touch the body part the robot is moving.

Once a behavior has been triggered, the drive is reset to its default value and decays again as explained above (the amount of the decay being reduced according to what has been acquired).

2) Drive to express knowledge: This drive regulates how the iCub expresses the acquired knowledge through synchronized speech, pointing and gaze. It aims at maintaining the interaction with the human by proactively informing her/him about its current state of knowledge. The drive level decays proportionally to the amount of already acquired information about the present entities. When below a given threshold (meaning that a significant amount of information has been acquired), it triggers a behavior alternating gazing toward the human and a known entity, synchronized with speech expressing the knowledge verbally, e.g. "This is the octopus", or "I know you, you are Daniel". Once such a behavior has been triggered, the drive is reset to its default value and decays again as explained above (the amount of the decay changing according to what is learned by satisfying the drive for knowledge acquisition).

These two drives allow the robot to balance knowledge acquisition and expression in an autonomous and dynamic way. At the beginning of the interaction, the robot does not know much about the current entities and therefore favors behaviors for knowledge acquisition. By acquiring more and more knowledge, it progressively switches to behaviors for knowledge expression. If new entities are introduced, e.g. a new object or another human, it will switch back to triggering more behaviors for knowledge acquisition and so on.

C. Goal-oriented behavior

The representations acquired by satisfying the drives above allow a more complex interaction through goal-oriented behavior managed by the contextual layer (see Figure 1 and Section III-D). Goals can be provided to the iCub from human speech and a meaning is extracted by the *language reservoir handler*, forming the representation of a goal in the *goals* module. Each goal is associated with an action plan on the form of a sequence of actions together with their pre- and post-conditions in the *action plans* module. The *action selection* module takes care of the execution of the plan according to the *associations* known to the robots, triggering the appropriate behaviors according to its current perception of the scene updated in real time. Goal achievement bypasses the reactive behavior described in the previous subsection by freezing all the drive levels during the execution of the plan. The available goals are described below.

1) Give or take an object: These goals are generated from a human verbal request, e.g. *"Give me the octopus"* or *"Take the cube"*. Here, the goal is represented as a region on the table, either the human area H (for the "Give" goal) or the iCub area I (for the "Take" goal), where the mentioned

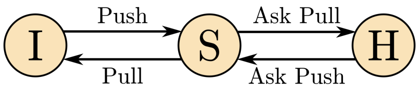

Figure 3. State transition graph used for generating the action plans of the goals "Give" and "Take". Nodes correspond to the table regions indicated in Figure 2. Arrows correspond to the actions to be executed for realizing a transition. According to the observed current state of the object, the iCub will execute the action which brings it closer to the goal state. For example, if the goal is to take an object which is in the human area, the robot will first ask the human to push it, then it will pull it.

object should be placed. Action plans are generated from the state-transition graph shown in Figure 3. State perception is updated in real-time according to the perceived location of the object computed through stereo-vision in the *object recognition* module.

2) Point to an object: This goal is generated through a verbal request, e.g. *"Point to the octopus"*. If the mentioned object is not known to the iCub, it will first ask the human to point to it in order to learn the new association between the name and the object visual representation. Once the name is known, or if it was already known, the iCub will point to the object.

3) Say the name of a recognized action: This goal is generated through a verbal request, e.g. *"How do you call this action?"* formulated just after the human has performed an action on an object. Four actions can be recognized: "push", "pull", "lift", "drop", "wave", and "point". The reply from the iCub provides the name of the action and the object as well as the hand used by the human, e.g. *"You pushed the cube with your left hand"*. The action classification is realized by the *synthetic sensory memory* module.

4) Tell what happened during a past interaction with a human: This goal is generated through a verbal request, e.g. *"What have you done the other day?"*. Based on its past interactions with the environment and with humans, the iCub has stored all the relevant information in its *autobiographical memory* (see Section III-D), including its own drives, motivations and actions, as well of the actions of the human (both, spoken and physically enacted). The *narrative handler* functional module is able to generate a narrative discourse from the content of the *autobiographical memory* and to generate an action plan on the form of a sequence of sentences. The human can request more detailed information about an event using sentences like *"What happened next?"* or *"Why did you do that?"*, this later question being answered according to the stored drive levels and goal states of the considered events. The robot can also learn a set of questions that it can re-use in the context of another story. Figure 4 shows an example of a generated narrative.

First **I wanted** *to* **get** *the* **toy**.

First **you have** *the* **toy**.

Then **I fail** *to* **grasp it**.

After **I fail** *to* **grasp**, **I reasoned**.

Because **I reasoned**, **I ask** *for the* **toy** *to* **you**

Finally **you gave me** *the* **toy**

Now **I have** *the* **toy**

I have *the* **toy** because **you gave it** *to* **me**

You gave *the* **toy** *to* **me** because **I wanted it**

Figure 4. Example of a narrative generated by the robot. The *language reservoir handler* will decompose the words in the narrative discourse in 3 categories: the discourse function words (DFW) which direct the discourse from one sentence to the other, the **open class words** (OCW) which correspond to the meaningful words in terms of vocabulary of the sentence, and the *closed class words* (CCW) which have a grammatical function in the sentence (see [70]).

5) Show the learned kinematic structure: As for the previous goals, this goal is generated through verbal requests. When asked "What have you learned from your arm babbling", the iCub refers the human to look at the screen where the kinematic structures of its arms are displayed. Also lines which connect nodes of the kinematic structures indicate the correspondences which the iCub has found between its left and right arm. Similarly, the iCub displays the correspondences which it has found between one of its arms and the body of the human (see Figure 5).

D. Scenario Progression

We propose to evaluate our system in a mixed-initiative scenario, where the iCub behaves autonomously as described in Section IV-B, and so does the human. The human can interrupt the robot behavior by formulating verbal requests as described in Section IV-C. The scenario can follow various paths according to the interaction between the iCub's internal drive dynamics, its perception of the environment and the behavior of the human.

We provide a series of videos of live interactions in a lab environment at http://clement-moulin-frier.github.io/dac-h3/.

Here, we describe one particular instance of the scenario. Figure 6 shows the corresponding drive dynamics and human-robot interactions, and Figure 1 shows the connections between the modules of the cognitive architecture. Each of the numbered items below is refers to its corresponding number in Figure 6.

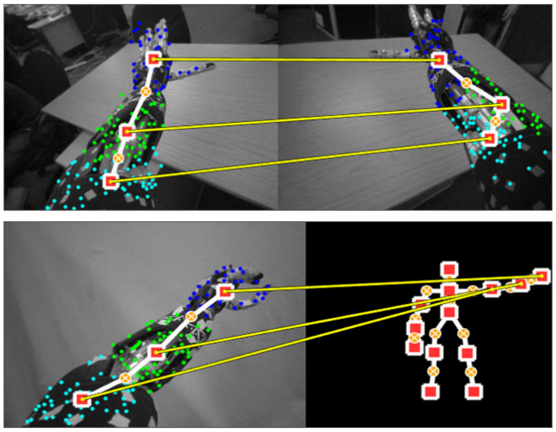

Figure 5. Examples of the kinematic structure correspondences which the iCub has found. The top figure shows the correspondences between the left and right arm of the iCub, which can be used to infer the body part names of one arm if the corresponding names of the other arm are known. Similarly, the bottom figure shows correspondences between the robot's body and the human's body.

1) At the beginning of the interaction, the iCub does not know much information about the current scene. In the *sensations* module, the *agent detector* detects the presence of a human and extracts its skeleton. The *object recognition* module performs blob detection for extracting objects on the table from the visual input of the eye cameras and trains a classifier to categorize them. The 3D locations of the present objects are also computed in *object recognition* through stereo vision. This form a first incomplete representation of the scene in the *episodic memory* where the *object property collector* registers the location and type of each detected entity (here objects and an agent). It also contain slots for unknown body parts of the iCub, here the five fingers of its right hand.
2) The presence of a large amount of missing information (presence of unknown objects, human and body parts) in the *sensations* module makes the drive for knowledge acquisition decaying rapidly in the *allostatic controller* and the drive for knowledge expression is kept constant (since there is no knowledge to express yet).
3) When this drive level is below threshold, it triggers the associated behavior (*behaviors* module) for acquiring information about an entity. The choice of the unknown entity is pseudo-random, with priority for requesting the name of an unknown human. This makes the robot look at the human. The visual input is passed to the *perceptions* module where the *synthetic sensory memory* segments the face from the background and attempts to recognize it from previously seen faces. It does not recognize the face and asks “I do not know you, who are you?”. The human replies, e.g., “I am Daniel”. The level of the drive is reset to its default value and both drives are frozen during the behavior execution.
4) The perceived speech is analyzed by the *language reservoir handler* in *perceptions* to extract the name “Daniel” and is associated with the face representation in the *associations* module. Thus, the next time the iCub will interact with this person, it will directly recognize him and not ask for his name.
5) Once this interaction is achieved, the drives continue to decay. Since the iCub has just acquired more information, the decay of the drive for knowledge acquisition is slower and the one for knowledge expression is increased. Still, the drive for knowledge acquisition reaches the threshold first. The behavior for acquiring information is therefore triggered again. This time, the random choice of an unknown entity makes the robot point to an object and ask “What is this object?”. The human replies e.g. “This is the cube”. The *language reservoir handler* extracts the name of the object from the reply and the *associations* module associates it with the visual representation of the pointed object from *object recognition*. Now the cube can later be referred by its name.
6) The drives continues to decay. This time, the drive for knowledge expression reaches the threshold first. This triggers the behavior for expressing the acquired knowledge. A known entity is chosen, here the cube, which is pointed while saying “This is a cube”.

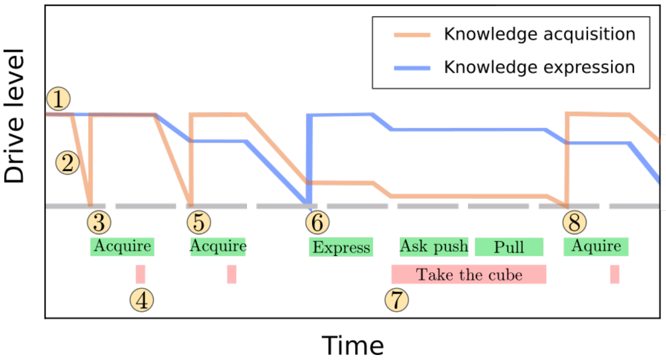

Figure 6. Drive level dynamics during a typical mixed-initiative scenario described in Section IV-D. Each drive starts at its default value and decays following the dynamics described in Section IV-B. When reaching a given threshold (dashed horizontal line) the associated behavior is triggered (green rectangles), the corresponding drive level is reset to its default value and both drive levels are frozen for the duration of the behavior. Human behavior is indicated by the red rectangles, being either a reply to a question asked by the iCub (small rectangles), or a request to the iCub triggering goal oriented behavior (here: “Take the cube”). The numbers refer to the description of the scenario progression in the main text.

7) The human asks “Take the cube”. A meaning is extracted by the *language reservoir handler* in *perceptions* and forms the representation of a goal to achieve in the *goal* module (here the desired location of the object, i.e. the region of the iCub I for the goal “take”, see Figure 3). An action plan is built in *action plans* with the sequence of two actions “Ask the human to push the object” then “Pull the object”, together with their pre- and post-conditions in term of states (I, S or H). The *action selection* module takes care of the realization of the plan. First, it instantiates the action plan toward the considered object, here the cube, through its connection with *associations*. Then, it executes each action until its associated post-condition is met (repeating it three times before giving up). Since the cube is in the human area H, the iCub first triggers the behavior for interacting verbally with the human, asking “Can you please bring the cube closer to the shared area?”. The human pushes the cube to the shared area S and the state transition is noticed by the robot thanks to the real-time object localization performed in *object recognition*. Then the robot triggers a motor action to pull the cube. Once the goal is achieved (i.e. the cube is in I), the drive levels which were frozen during this interaction continue to decay.
8) The drive for knowledge acquisition reaches the threshold first. The associated behavior now chooses to acquire the name of a body part. The robot triggers the behavior for raising its hand and moving a random unknown body part, here the middle finger. It looks at the human and asks “How do you call this part of my body?”. The name of the body part is extracted from the human’s reply and is associated with the joint that was moved in *associations*.

The interaction continues following the dynamics of the drives and interrupted by the requests from the human. Once all available information about the present entities is acquired, the drive for knowledge acquisition stops to decay. However, the robot still maintains the interaction through its drive for knowledge expression and the human can still formulate requests for goal-oriented behavior. When new entities are introduced, e.g. an unknown object or another human entering in the scene, the drive for knowledge acquisition decays again and the process continues.

V. Conclusion and Future Works

This paper has introduced *DAC-h3*, a proactive robot cognitive architecture to acquire and express knowledge about the world and the self. The architecture is based on the Distributed Adaptive Control (DAC) theory of the brain and mind, which provides a biologically grounded framework for organizing various functional modules into a coherent cognitive architecture. Those modules implement state-of-the-art algorithms modeling various cognitive functions for autonomous self-regulation, whole-body motor control, multimodal perception, knowledge representation, natural language processing, and goal-oriented behavior. They are all implemented using the YARP robotic middleware on the iCub robot, ensuring their parallel execution in real time and providing synchronous and asynchronous communication protocols among modules.

The main contribution of this paper is not about the modeling of the specific functional modules, which already have been published (see Section III), but rather about the integration of a heterogeneous collection of modules into a coherent and operational cognitive architecture. For this aim, the *DAC-h3* architecture is organized as a layered control structure with tight coupling within and between layers (Figure 1 and Section III): the *somatic*, *reactive*, *adaptive*, and *contextual* layers. Across these layers, a columnar organization exists that deals with the processing of states of the world or exteroception, the self or interoception and action. Two main control loops generate the behavior of the robot. First, a reactive-adaptive control loop ensures autonomy and proactivity through the self-regulation of internal drives for knowledge acquisition and expression. It allows the robot to proactively manage its own knowledge acquisition process and to maintain the interaction with a human partner, while associating multimodal information about entities with their linguistic labels. Second, an adaptive-contextual control loop allows the robot to satisfy human requests, triggering goal-oriented behavior relying on the acquired knowledge. Those goal-oriented behaviors are related to action planning for object passing, pointing, action recognition, narrative expression and kinematic structure learning demonstration.

We have implemented the entire *DAC-h3* architecture and presented a HRI scenario where an iCub humanoid robot interacts with objects and a human to acquire information about the present objects and agents as well as its own body parts. We have analyzed a typical interaction in detail, showing how *DAC-h3* is able to dynamically balance the knowledge acquisition and expression processes according to the properties of the environment, and to deal with a mixed initiative scenario where both the robot and the human are behaving autonomously. In a series of video recordings, we show the ability of *DAC-h3* to adapt to different situations and environments.

The current work has the following limitations. First, some of the available abilities deserve to be better integrated into the HRI scenario. For example, this is the case for the kinematic structure learning process which is currently executed in a separated learning phase instead of being fully integrated within the interaction scenario. Similarly, the narrative can only be generated from specific chunks of the robot's history as recorded in the autobiographical memory. Second, in this paper we do not provide a statistical analysis of the HRI experiments. The reason is that we focus on the description of the entire architecture and on their theoretical principles. Statistical analysis will allow to better demonstrate the utility of some of these principles, for example how proactivity helps solving the referential indeterminacy problem, as well as the effect of the robot's autonomy on the acceptability of the system by naive users. Third, although *DAC-h3* is able to solve parts of the *Symbol Grounding Problem* (SGP), it still presupposes a symbolic concept of entity which is given *a priori* to the system. Therefore, our contribution is more about the ability to proactively acquire multimodal information about these entities, and linking them to linguistic labels that can be reused to express complex goal-oriented behavior later on.

We are currently extending the proposed architecture in the following ways. First, we are better integrating some of the available abilities within the interaction scenario as mentioned above. This will allow to start the knowledge acquisition process from scratch in a fully autonomous way. Second, we are considering to use more biologically plausible models for some of the existing modules, namely the *action planning* and *action selection* modules. These are currently algorithmic implementations using predefined action plans. We want to replace it with an existing model of rule learning grounded in the neurobiology of the prefrontal cortex which is able to learn optimal action policies from experience to maximize long-term reward [**?**]. An interesting feature of this model for solving the SGP is that it relies on neural memory-units encoding sensorimotor contingencies with causal relationships learned through adaptive connections between them. Such sensorimotor contingencies, adaptively learned from experience, could provide the basis of emerging symbols without relying on a pre-existing concept of entity as in the current version of the architecture. Third, we are also integrating more low-level reactive control abilities through an acquired notion of a peripersonal space [**?**], where the robot will be able to optimize its own action primitives to maintain safety distances with aversive objects (e.g. a spider) in real time while executing reaching actions toward other objects. Finally, we are working on a self-exploration process to autonomously discover the area which is reachable by the robot, and subsequently employing this self-model and applying it to the human partner to estimate his/her reachability.

Acknowledgment

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. FP7-ICT-612139 (What You Say Is What You Did project).

References

[1] S. Harnad, "The symbol grounding problem," *Physica D: Nonlinear Phenomena*, vol. 42, no. 1, pp. 335 – 346, 1990.

[2] S. Coradeschi and A. Saffiotti, "An introduction to the anchoring problem," *Robotics and Autonomous Systems*, vol. 43, no. 2–3, pp. 85 – 96, 2003.

[3] T. Taniguchi, T. Nagai, T. Nakamura, N. Iwahashi, T. Ogata, and H. Asoh, "Symbol emergence in robotics: a survey," *Advanced Robotics*, vol. 30, no. 11-12, pp. 706–728, 2016.

[4] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. Von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor *et al.*, "The iCub humanoid robot: An open-systems platform for research in cognitive development," *Neural Networks*, vol. 23, no. 8, pp. 1125–1134, 2010.

[5] S. Lallée, K. Hamann, J. Steinwender, F. Warneken, U. Martienz, H. Barron-Gonzales, U. Pattacini, I. Gori, M. Petit, G. Metta *et al.*, "Cooperative human robot interaction systems: IV. Communication of shared plans with Naïve humans using gaze and speech," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 129–136.

[6] S. Lallée, U. Pattacini, S. Lemaignan, A. Lenz, C. Melhuish, L. Natale, S. Skachek, K. Hamann, J. Steinwender, E. A. Sisbot *et al.*, "Towards a platform-independent cooperative human robot interaction system: III an architecture for learning and executing actions and shared plans," *IEEE Transactions on Autonomous Mental Development*, vol. 4, no. 3, pp. 239–253, 2012.

[7] M. Petit, S. Lallée, J.-D. Boucher, G. Pointeau, P. Cheminade, D. Ognibene, E. Chinellato, U. Pattacini, I. Gori, U. Martinez-Hernandez *et al.*, "The coordinating role of language in real-time multimodal learning of cooperative tasks," *IEEE Transactions on Autonomous Mental Development*, vol. 5, no. 1, pp. 3–17, 2013.

[8] S. Calinon, F. D'halluin, E. L. Sauser, D. G. Caldwell, and A. G. Billard, "Learning and reproduction of gestures by imitation," *IEEE Robotics & Automation Magazine*, vol. 17, no. 2, pp. 44–54, 2010.

[9] Y. Demiris and A. Meltzoff, "The robot in the crib: A developmental analysis of imitation skills in infants and robots," *Infant and Child Development*, vol. 17, no. 1, pp. 43–53, 2008.

[10] S. Schaal, "Is imitation learning the route to humanoid robots?" *Trends in cognitive sciences*, vol. 3, no. 6, pp. 233–242, 1999.

[11] M. Ewerton, G. Maeda, J. Peters, and G. Neumann, "Learning motor skills from partially observed movements executed at different speeds," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015, pp. 456–463.

[12] M. Quine, *Word and object*. MIT Press, 1960.

[13] P. Vogt and F. Divina, "Social symbol grounding and language evolution," *Interaction Studies*, vol. 8, no. 1, pp. 31–52, 2007.

[14] A. Cangelosi, "The grounding and sharing of symbols," *Pragmatics & Cognition*, vol. 14, no. 2, pp. 275–285, 2006.

[15] S. C. Levinson, "On the human interaction engine," in *Wenner-Gren Foundation for Anthropological Research, Symposium 134*. Berg, 2006, pp. 39–69.

[16] U. Liszkowski, M. Carpenter, A. Henning, T. Striano, and M. Tomasello, "Twelve-month-olds point to share attention and interest," *Developmental science*, vol. 7, no. 3, pp. 297–307, 2004.

[17] D. E. Berlyne, "A theory of human curiosity," *British Journal of Psychology. General Section*, vol. 45, no. 3, pp. 180–191, 1954.

[18] M. Tomasello, M. Carpenter, J. Call, T. Behne, and H. Moll, "Understanding and sharing intentions: The origins of cultural cognition," *Behavioral and brain sciences*, vol. 28, no. 05, pp. 675–691, 2005.

[19] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3, pp. 143–166, 2003.

[20] J. Pineau, M. Montemerlo, M. Pollack, N. Roy, and S. Thrun, "Towards robotic assistants in nursing homes: Challenges and results," *Robotics and autonomous systems*, vol. 42, no. 3, pp. 271–281, 2003.

[21] I. R. Nourbakhsh, J. Bobenage, S. Grange, R. Lutz, R. Meyer, and A. Soto, "An affective mobile robot educator with a full-time job," *Artificial Intelligence*, vol. 114, no. 1, pp. 95–124, 1999.

[22] A. G. Di Nuovo, D. Marocco, S. Di Nuovo, and A. Cangelosi, "Autonomous learning in humanoid robotics through mental imagery," *Neural Networks*, vol. 41, pp. 147–155, 2013.

[23] B. Adams, C. Breazeal, R. A. Brooks, and B. Scassellati, "Humanoid robots: a new kind of tool," *IEEE Intelligent Systems and their Applications*, vol. 15, no. 4, pp. 25–31, 2000.

[24] K. Dautenhahn and A. Billard, "Bringing up robots or—the psychology of socially intelligent robots: From theory to implementation," in *Proceedings of the Conference on Autonomous Agents*, 1999, pp. 366–367.

[25] A. Newell, *Unified theories of cognition*. Harvard University Press, 1990.

[26] A. Newell, J. C. Shaw, and H. A. Simon, "Report on a general problem-solving program," *IFIP Congress*, pp. 256–264, 1959.

[27] J. E. Laird, A. Newell, and P. S. Rosenbloom, "SOAR: An architecture for general intelligence," *Artificial Intelligence*, vol. 33, no. 1, pp. 1–64, 1987.

[28] J. R. Anderson, *The Architecture of Cognition*, M. Cambridge, Ed. Harvard University Press, 1983.

[29] R. A. Brooks, "Intelligence without representation," *Artificial Intelligence*, vol. 47, no. 1-3, pp. 139–159, 1991.

[30] R. Brooks, "A robust layered control system for a mobile robot," *IEEE Journal on Robotics and Automation*, vol. 2, no. 1, pp. 14–23, 1986.

[31] Á. Miklósi and M. Gácsi, "On the utilization of social animals as a model for social robotics," *Frontiers in Psychology*, vol. 3, no. 75, pp. 1–10, 2012.

[32] K. Dautenhahn, "Socially intelligent robots: dimensions of human-robot interaction." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, vol. 362, no. 1480, pp. 679–704, 2007.

[33] P. Carruthers and P. K. Smith, "Theories of Theories of Mind," *Australasian Journal of Philosophy*, vol. 76, 1998.

[34] E. Di Paolo and H. De Jaegher, "The interactive brain hypothesis," *Frontiers in Human Neuroscience*, vol. 6, no. 163, pp. 1–16, 2012.

[35] T. Taniguchi, "Symbol emergence in robotics for long-term human-robot collaboration," in *IFAC Symposium on Analysis, Design, and Evaluation of Human-Machine Systems*, 2016, pp. 144 – 149.

[36] B. Scassellati, "Investigating models of social development using a humanoid robot," in *Proceedings of the IEEE International Joint Conference on Neural Networks*, 2003, pp. 2704–2709.

[37] A. Stoytchev and R. Arkin, "Combining deliberation, reactivity, and motivation in the context of a behavior-based robot architecture," in *Proceedings of the IEEE International Symposium on Computational Intelligence in Robotics and Automation*, 2001, pp. 290–295.

[38] M. Malfaz, Á. Castro-Gonzalez, R. Barber, and M. A. Salichs, "A Biologically Inspired Architecture for an Autonomous and Social Robot," *IEEE Transactions on Autonomous Mental Development*, vol. 3, no. 3, pp. 232–246, 2011.

[39] P. F. M. J. Verschure, T. Voegtlin, and R. J. Douglas, "Environmentally mediated synergy between perception and behaviour in mobile robots," *Nature*, vol. 425, no. 6958, pp. 620–624, 2003.

[40] P. F. M. J. Verschure, C. M. A. Pennartz, and G. Pezzulo, "The why, what, where, when and how of goal-directed choice: neuronal and computational principles," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1655, pp. 1–14, 2014.

[41] J.-Y. Puigbò, C. Moulin-Frier, and P. F. Verschure, "Towards self-controlled robots through distributed adaptive control," in *Biomimetic and Biohybrid Systems*, 2016, pp. 490–497.

[42] L. Steels, "The synthetic modeling of language origins," *Evolution of Communication*, vol. 1, no. 1, pp. 1–34, 1997.

[43] F. Kaplan, "Semiotic schemata: Selection units for linguistic cultural evolution," in *Proceedings of the International Conference on Artificial Life*, 2000, pp. 372–381.

[44] C. Moulin-Frier, J. Diard, J.-L. Schwartz, and P. Bessière, "COSMO ("Communicating about Objects using Sensory-Motor Operations"): a Bayesian modeling framework for studying speech communication and the emergence of phonological systems," *Journal of Phonetics*, vol. 53, pp. 5–41, 2015.

[45] A. Billard and K. Dautenhahn, "Grounding communication in autonomous robots: an experimental study," *Robotics and Autonomous Systems*, vol. 24, no. 1, pp. 71–79, 1998.

[46] A. Cangelosi, E. Hourdakis, and V. Tikhanoff, "Language acquisition and symbol grounding transfer with neural networks and cognitive robots," in *IEEE International Joint Conference on Neural Network Proceedings*, 2006, pp. 1576–1582.

[47] D. Marocco, A. Cangelosi, K. Fischer, and T. Belpaeme, "Grounding action words in the sensorimotor interaction with the world: experiments with a simulated iCub humanoid robot," *Frontiers in Neurorobotics*, vol. 4, no. 7, pp. 1–15, 2010.

[48] S. Ivaldi, S. Nguyen, N. Lyubova, A. Droniou, V. Padois, D. Filliat, P.-Y. Oudeyer, and O. Sigaud, "Object learning through active exploration," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 1, pp. 56–72, 2013.

[49] P.-Y. Oudeyer, F. Kaplan, and V. V. Hafner, "Intrinsic motivation systems for autonomous mental development," *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 265–286, 2007.

[50] C. Breazeal and B. Scassellati, "Infant-like social interactions between a robot and a human caregiver," *Adaptive Behavior*, vol. 8, no. 1, pp. 49–74, 2000.

[51] O. C. Schrempf, U. D. Hanebeck, A. J. Schmid, and H. Worn, "A novel approach to proactive human-robot cooperation," in *IEEE International Workshop on Robot and Human Interactive Communication*, 2005, pp. 555–560.

[52] I. Lutkebohle, J. Peltason, L. Schillingmann, B. Wrede, S. Wachsmuth, C. Elbrechter, and R. Haschke, "The curious robot - structuring interactive robot learning," in *IEEE International Conference on Robotics and Automation*, 2009, pp. 4156–4162.

[53] F. Broz, C. L. Nehaniv, T. Belpaeme, A. Bisio, K. Dautenhahn, L. Fadiga, T. Ferrauto, K. Fischer, F. Förster, O. Gigliotta *et al.*, "The italk project: A developmental robotics approach to the study of individual, social, and linguistic learning," *Topics in cognitive science*, vol. 6, no. 3, pp. 534–544, 2014.

[54] A. Cangelosi *et al.*, "Integration of action and language knowledge: A roadmap for developmental robotics," *IEEE Transactions on Autonomous Mental Development*, vol. 2, no. 3, pp. 167–195, 2010.

[55] A. Antunes, L. Jamone, G. Saponaro, A. Bernardino, and R. Ventura, "From human instructions to robot actions: Formulation of goals, affordances and probabilistic planning," in *IEEE International Conference on Robotics and Automation*, 2016, pp. 5449–5454.

[56] E. A. Krause, M. Zillich, T. E. Williams, and M. Scheutz, "Learning to Recognize Novel Objects in One Shot through Human-Robot Interactions in Natural Language Dialogues," in *AAAI Conference on Artificial Intelligence*, 2014, pp. 2796–2802.

[57] J. Dias, W. C. Ho, T. Vogt, N. Beeckman, A. Paiva, and E. André, "I know what I did last summer: Autobiographic memory in synthetic characters," in *International Conference on Affective Computing and Intelligent Interaction*, 2007, pp. 606–617.

[58] G. Sieber and B. Krenn, "Towards an episodic memory for companion dialogue," in *International Conference on Intelligent Virtual Agents*, J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, and A. Safonova, Eds., 2010, pp. 322–328.

[59] D. S. Syrdal, K. Dautenhahn, K. L. Koay, and W. C. Ho, "Views from within a narrative: Evaluating long-term human–robot interaction in a naturalistic environment using open-ended scenarios," *Cognitive Computation*, vol. 6, no. 4, pp. 741–759, 2014.

[60] P. Fitzpatrick, G. Metta, and L. Natale, "Towards long-lived robot genes," *Robotics and Autonomous systems*, vol. 56, no. 1, pp. 29–45, 2008.

[61] M. Sanchez-Fibla, U. Bernardet, E. Wasserman, T. Pelc, M. Mintz, J. C. Jackson, C. Lansink, C. Pennartz, and P. F. Verschure, "Allostatic control for robot behavior regulation: a comparative rodent-robot study," *Advances in Complex Systems*, vol. 13, no. 3, pp. 377–403, 2010.

[62] M. S. Fibla, U. Bernardet, and P. F. Verschure, "Allostatic control for robot behaviour regulation: An extension to path planning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 1935–1942.

[63] U. Pattacini, F. Nori, L. Natale, and G. Metta, "An experimental evaluation of a novel minimum-jerk cartesian controller for humanoid robots," in *The IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 1668–1674.

[64] A. Parmiggiani, M. Maggiali, L. Natale, F. Nori, A. Schmitz, N. Tsagarakis, J. S. Victor, F. Becci, G. Sandini, and G. Metta, "The Design of the iCub Humanoid Robot," *International Journal of Humanoid Robotics*, vol. 9, no. 4, 2012.

[65] S. Lallee and P. F. Verschure, "How? Why? What? Where? When? Who? Grounding Ontology in the Actions of a Situated Social Agent," *Robotics*, vol. 4, no. 2, pp. 169–193, 2015.

[66] V. Vouloutsi, K. Grechuta, S. Lallée, and P. F. Verschure, "The influence of behavioral complexity on robot perception," in *Conference on Biomimetic and Biohybrid Systems*, 2014, pp. 332–343.

[67] G. Pasquale, C. Ciliberto, L. Rosasco, and L. Natale, "Object identification from few examples by improving the invariance of a deep convolutional neural network," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016.

[68] X. Hinaut and P. F. Dominey, "Real-time parallel processing of grammatical structure in the fronto-striatal system: a recurrent network simulation study using reservoir computing," *PloS one*, vol. 8, no. 2, pp. 1–18, 2013.

[69] X. Hinaut, M. Petit, G. Pointeau, and P. F. Dominey, "Exploring the acquisition and production of grammatical constructions through human-robot interaction with echo state networks," *Frontiers in Neurorobotics*, vol. 8, no. 16, pp. 1–17, 2015.

[70] A.-L. Mealier, G. Pointeau, P. Gardenfors, and P.-F. Dominey, "Construals of meaning: The role of attention in robotic language production," *Interaction Studies*, vol. 17, no. 1, pp. 41–69, 2016.

[71] A. E. Goldberg, *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, 1995.

[72] M. Zambelli, T. Fischer, M. Petit, H. J. Chang, A. Cully, and Y. Demiris, "Towards anchoring self-learned representations to those of other agents," in *Workshop on Bio-inspired Social robot Learning in Home Scenarios at IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016.

[73] M. Zambelli and Y. Demiris, "Online Multimodal Ensemble Learning using Self-learnt Sensorimotor Representations," in *IEEE Transactions on Cognitive and Developmental Systems*, 2016 (accepted).

[74] H. Jaeger, "Adaptive nonlinear system identification with echo state networks," in *Advances in Neural Information Processing Systems*, 2002, pp. 593–600.

[75] H. Soh and Y. Demiris, "Spatio-temporal learning with the Online Finite and Infinite Echo-state Gaussian processes," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 3, pp. 522–536, 2015.

[76] S. Vijayakumar and S. Schaal, "Fast and efficient incremental learning for high-dimensional movement systems," in *IEEE International Conference on Robotics and Automation*, 2000, pp. 1894–1899.

[77] H. J. Chang and Y. Demiris, "Unsupervised learning of complex articulated kinematic structures combining motion and skeleton information," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3138–3146.

[78] H. J. Chang, T. Fischer, M. Petit, M. Zambelli, and Y. Demiris, "Kinematic structure correspondences via hypergraph matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4216–4225.

[79] T. Fischer and Y. Demiris, "Markerless Perspective Taking for Humanoid Robots in Unconstrained Environments," in *IEEE International Conference on Robotics and Automation*, 2016, pp. 3309–3316.

[80] P. Michelon and J. M. Zacks, "Two kinds of visual perspective taking," *Perception & Psychophysics*, vol. 68, no. 2, pp. 327–337, 2006.

[81] T. J. Prescott, N. Lepora, and P. F. M. J. Vershure, "A future of living machines?: International trends and prospects in biomimetic and biohybrid systems," in *Proceedings of the SPIE 9055, Bioinspiration, Biometrics and Bioreplication*, 2014.

[82] P. Verschure, "Formal minds and biological brains II: From the mirage of intelligence to a science and engineering of consciousness," in *IEEE Intelligent Systems Trends and Controversies*, 2013, pp. 33–36.

[83] G. Pointeau, M. Petit, and P. F. Dominey, "Successive Developmental Levels of Autobiographical Memory for Learning Through Social Interaction," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 3, pp. 200–212, Sep. 2014.

[84] M. Petit, T. Fischer, and Y. Demiris, "Lifelong Augmentation of Multi-Modal Streaming Autobiographical Memories," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 8, no. 3, pp. 201–213, 2016.

[85] M. Petit, T. Fischer, and Y. Demiris, "Towards the emergence of procedural memories from lifelong multi-modal streaming memories for cognitive robots," in *Workshop on Machine Learning Methods for High-Level Cognitive Capabilities in Robotics at IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2016.

[86] A. Damianou, C. H. Ek, L. Boorman, N. D. Lawrence, and T. J. Prescott, "A top-down approach for a synthetic autobiographical memory system," in *Biomimetic and Biohybrid Systems*, 2015, pp. 280–292.

[87] A. C. Damianou and N. D. Lawrence, "Deep gaussian processes." in *Artificial Intelligence and Statistics Conference*, 2013, pp. 207–215.

[88] D. Camilleri, A. Damianou, H. Jackson, N. Lawrence, and T. Prescott, "iCub Visual Memory Inspector: Visualising the iCub's Thoughts," in *Conference on Biomimetic and Biohybrid Systems*, 2016, pp. 48–57.