AFRL-HE-WP-TR-2007-0090

An Integrated Architecture for Grounded Intelligence in its Development, Experiential, Environmental, and Social Context

Cynthia Breazeal
Larry Barsalou
Linda Smith

MIT Media Laboratory
Massachusetts Institute of Technology
20 Ames Street
Cambridge MA 02139

May 2007

Final Report for August 2005 to May 2007

Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Cognitive Systems Branch
Wright-Patterson AFB OH 45433

NOTICE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory, Det 1, Wright Site, Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

THIS REPORT HAS BEEN REVIEWED AND IS APPROVED F OR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

AFRL-HE-WP-TR-2007-0090

//SIGNED//

JOHN L. CAMP
Work Unit Manager
Cognitive Systems Branch

//SIGNED//

DANIEL G. GODDARD
Chief, Warfighter Interface Division
Human Effectiveness Directorate
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. REPORT DATE (DD-MM-YYYY)	2. REPORT TYPE	3. DATES COVERED (From - To)
May 2007	Final	August 2005 - May 2007

4. TITLE AND SUBTITLE
An Integrated Architecture for Grounded Intelligence in its Development, Experiential, Environmental, and Social Context

5a. CONTRACT NUMBER
FA8650-05-C-7255

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
63123F

6. AUTHOR(S)
Cynthia Breazeal, Larry Barsalou, Linda Smith

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER
DRPA0510

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
MIT Media Laboratory
Massachusetts Institute of Technology
20 Ames Street
Cambridge MA 02139

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Materiel Command
Air Force Research Laboratory
Human Effectiveness Directorate
Warfighter Interface Division
Cognitive Systems Branch
Wright-Patterson AFB OH 45433-7604

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/HECS

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-HE-WP-TR-2007-0090

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.
AFRL/PA cleared on 13 September 2007, AFRL-WS-07-2101.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This final report describes a novel biologically-inspired cognitive architecture. The systems level architecture described in this report was inspired by Simmons and Barsalou (2003), with psychologically-inspired additions proposed by Breazeal et al (2005), augmented by key insights from developmental psychology (Smith, 2005). A cognitive agent will have multiple modality-specific systems, including sensory systems (e.g., vision, audition, touch), a motor system, an emotional system, and a cognitive system. As a system perceives its external world and internal mental states, feature systems will represent these experiences in the relevant perceptual modalities, and a hierarchical system of neurally-inspired association areas will capture them, so that they can be reenacted or simulated in the future. These perceptually grounded representations guide the intelligent and social operations of agents.

15. SUBJECT TERMS Cognitive Architecture, Collaboration, Perception, Sense Making, Communication, Human Performance, Modeling and Simulation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			John L. Camp
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR	24	19b. TELEPHONE NUMBER (include area code)

THIS PAGE LEFT INTENTIONALLY BLANK

CONTENTS

LIST OF FIGURES

THIS PAGE LEFT INTENTIONALLY BLANK

1 Introduction

This final report describes a novel biologically-inspired cognitive architecture that results from many discussions and collaborative efforts by members of this team as well as fruitful discussions with members of other teams participating in the BICA Phase I program. As part of our Phase I efforts, we contributed ideas to the TOSCA architecture, which was the basis of our Phase II proposal.

The systems level architecture described in this final report is inspired by Simmons and Barsalou (2003), with psychologically-inspired additions proposed by Breazeal et al (2005), augmented by key insights from developmental psychology (Smith, 2005). A cognitive agent will have multiple modality-specific systems, including sensory systems (e.g., vision, audition, touch), a motor system, an emotional system, and a cognitive system. As a system perceives its external world and internal mental states, feature systems will represent these experiences in the relevant perceptual modalities, and a hierarchical system of neurally-inspired association areas will capture them, so that they can be reenacted or simulated in the future. These perceptually grounded representations guide the intelligent and social operations of agents.

Eight themes underlie our proposed system, derived from what is currently understood about the origin and developmental trajectory of the requisite cognitive skills in humans and brain science:

- The system operates here and now in the physical world.
- The system is internally motivated to act and learn (e.g., drives and affect), as well as externally motivated through social interaction with con-specifics
- The system also operates in the social world, working with other agents, human and robotic
- The system understands other agents in social terms (e.g., mind-reading, joint attention, perspective taking, etc.)
- The systems representations are grounded in modality-specific systems (e.g., vision, motor, emotion, etc.), in its body, and in its meta-cognition.
- The system changes as a consequence of its experience, exhibiting both short-term adaptation to experience and long-term developmental change. Imitation and other forms of social learning are particularly important.
- The systems overall architecture is neurally inspired at the systems level, based on the brains modality-specific systems and association areas.
- Although the system is built from a network of distributed computational processes (e.g., neural nets, agent-based architectures, etc.) and modality-specific representations, it nevertheless performs classic symbolic operations that appear central to human intelligence, such as predication and conceptual combination.

Our goal is to break new ground in designing an intelligent system whose cognition is grounded in the brains modality-specific system and shaped by environmental, social, and developmental constraints. Our approach is deeply informed by lessons from biology about how to build intelligence. This endeavor not only considers what is in the head as informed by neuroscience and psychology, but also considers the deep involvement of the rest of the body, as well as the nature of the environment the intelligence is situated within, and the activities in which it is engaged.

2 Theoretical Background

Embodied Cognition

During its first few decades of existence, artificial intelligence has not only *drawn* from theories of cognitive psychology, but to a wide extent also *shaped* notions of amodal, symbolic information processing. According to this view, information is translated from perceptual stimuli into nonperceptual symbols, later used for information retrieval, decision making, and action production. This view also corresponds to much of the work currently done in robotics, where sensory input is translated into semantic symbols, which are then operated upon for the production of motor control.

An increasing body of recent findings challenges this view and suggests instead that concepts and memory are phenomena grounded in modal representations utilizing many of the same mechanisms used during the perceptual process. A prominent theory explaining these findings is one of "simulators", siting memory and recall in the very neural modules that govern perception itself, subsequently used by ways of "simulation" or "imagery" [Barsalou, 1999, Kosslyn, 1995]. Perceptual symbols are organized in cross-modal networks of activation which are used to dynamically reconstruct and produce knowledge, concepts, and decision making. This view is supported by evidence of inter-modal behavioral influences, as well as by the detection of perceptual neural activation when a subject is using a certain concept in a non-perceptual manner (e.g. [Martin, 2001]).

Thus, when memory or language are invoked to produce behavior, the underlying perceptual processes elicit many of the same neural patterns and behaviors normally used to regulate perception [Spivey et al., 2005]. To name but a few examples, it has been shown that reading a sentence that has an implied orientation reduces response time on image recognition that is similarly oriented [Stanfield and Zwaan, 2001]; memory recall impairment is found to match speech impediments in children (mistaking rings for wings in children that pronounce 'r's as 'w's) [Locke and Kutz, 1975]; and comparing visually similar variations of a word (e.g. a pony's mane and a horse's mane) is faster than visually distinct variations (e.g. a lion's mane) [Barsalou et al., 1999].

In parallel to a perception-based theory of cognition lies an understanding that cognitive processes are equally interwoven with motor activity. Evidence in human developmental psychology shows that motor and cognitive development are not parallel but

highly interdependent. For example, research showed that artificially enhancing 3-month old infant's grasping abilities (through the wearing of a sticky mitten), equated some of their cognitive capabilities to the level of older, already grasping[1], infants [Somerville et al., 2004].

A related case has been made with regard to hand signals, which are viewed by some as foremost instrumental to lexical lookup during language generation [Krauss et al., 1996], and is supported by findings of redundancy in head-movements [McClave, 2000] and facial expression [Chovil, 1992] during speech generation.

A large body of work points to an isomorphic representation between perception and action, leading to mutual and often involuntary influence between the two [Wilson, 2001]. Some researches speak of specific 'privileged loops' from perception to action, for example between speech and auditory perception, or visual perception and certain motor activity, indicating these action systems' roles in the perceptual pathway [McLeod and Posner, 1984].

In addition, neurological findings indicate that—in some primates—observing an action and performing it causes activation in the same cerebral areas, a phenomenon labeled "mirror neurons" [Gallese and Goldman, 1996]. Similarly, listening to speech has been shown to activate motor area related to speech production [Wilson et al., 2004], and in pianists, listening to a tonal sequence triggers neural activation in areas associate with finger movement — both for the current tone, and for the next tone in cases where the subject is familiar with the melody. This common coding is thought to play a role in imitation, and the relation of the behavior of others to our own, which is considered a central process in the development of a Theory of Mind [Meltzoff and Moore, 1997], and possibly underlie the rapid and effortless adaptation to a partner that is needed to perform a joint task [Sebanz et al., 2006]. For a thorough review of these mirror neurological phenomena and the related connection between perception and action, as it pertains to human-robot interaction, see [Matarić, 2002].

An important insight of the evidence is that perceptual processing is not a strictly bottom-up analysis of raw available data, as it is often modeled in robotic systems. Instead, simulations of perceptual processes prime the acquisition of new perceptual data, motor knowledge is used in sensory parsing, and intentions, goals, and expectations all play a role in the ability to parse the world into meaningful objects. This seems to be particularly true for the parsing of human behavior in a goal-oriented and anticipatory manner, a vital component of joint action.

Experimental data supports this hypothesis, finding perception to be predictive (for a review, see [Wilson and Knoblich, 2005]). In vision, information is sent both upstream and downstream, and object priming triggers top-down processing, biasing lower-level mechanisms in sensitivity and criterion [Kosslyn, 1995]. Similarly, visual lip-reading affects the perception of auditory syllables indicating that the sound signal is not processed as a

[1] In the physical sense.

raw unknown piece of data [Massaro and Cohen, 1983][2]. High-level visual processing is also involved in the perception of human figures from point light displays, enabling subjects to identify gender and identity from very sparse visual information [Thornton et al., 1998].

While the idea of top-down processing has been utilized in some form in past object recognition systems, e.g. [Bregler, 1997, Hamdan et al., 1999] and in an amodal setting in collaborative planning systems, e.g. [Rich et al., 2001], the application of an integrated top-down implementation in a robotic behavior system is yet to be tested.

We will use the term "embodied cognition" to relate to the effect and interrelation of the above mentioned elements: (a) perceptual symbol systems, (b) integration between perception and action, and (c) top-down processing. We use this overarching term to denote an approach that views mental processes not as amodal semantic symbol processors with perceptual inputs and motor outputs, but as integrated psycho-physical systems acting as indivisible wholes.

3 Multimodal Representation via Simulators

This section describes the main components of the proposed research architecture, which stands on three main pillars, based on the theoretical framework outlined above: (a) perception-based memory; (b) action-perception activation networks; and (c) a motivation- and goal-based action selection mechanism. These three concepts are highly interrelated: the perceptual memory model supports top-down simulator- and emulator-type biasing, which is activated — among others — by the action-perception activation network connections. The motivation/goal layer both builds on the perception-action layer, by serving as a recourse for unresolved goals, and conversely triggers for simulator- and emulator-type biasing as part of the action selection mechanism.

Perception-Based Memory

Grounded in perceptual symbol theories of cognition, and in particular that of simulators [Barsalou, 1999], memories and concepts are based on perceptual snapshots and reside in the various perceptual systems rather than in an amodal centralized semantic network (Figure 2). Incoming perceptions are organized in percept trees [Blumberg et al., 2002] for each modality, organizing them on a gradient of specificity (Figure 1).

Memories are organized in activation networks, connecting them to each other, as well as to concepts, actions, plans, and intentions (the dotted lines in Figure 2). However, it

[2] As cited in [Wilson, 2001].

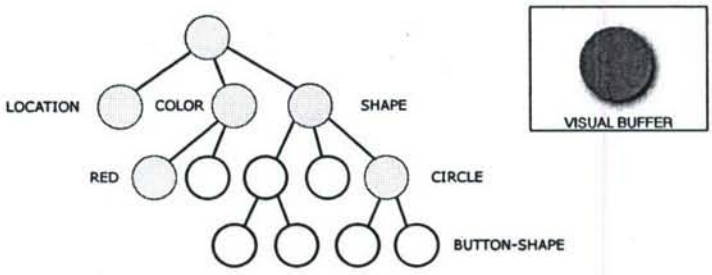

Figure 1: Percept trees organize sensory experience on a gradient of specificity.

is important to note that connections between perceptual, conceptual, and action memory are not binary. Instead, a single perceptual memory—for example the sound of a cat meowing—can be more strongly connected to one memory—like the proprioceptive memory of tugging at a cat's tail—or more weakly connected to another memory, such as that of one's childhood living room scent.

Perceptions are retained in memory, but decay over time. Instead of subscribing to the classic division of short- and long-term memory, the proposed architecture advocates a gradient information-based decay in which more specific perceptual memory decays faster, and more general perceptual information is retained longer. Decay is governed by the amount of storage needed to keep perceptual memory at various levels of specificity, retaining more specific memories for a shorter period of time, and more abstract—or filtered—perceptions longer.

These retentions are also influenced by the relevance of the perceptual information to the current task and their attentive saliency: perceptual elements that are specifically attended to at time of acquisition will be retained longer than ones that were peripheral at that time. In a similar vein, affective states of the system also influence memory retention[3].

For example in the visual processing pathway, raw images are retained for a short period of time, while colors, line orientations, blob location, and fully recognized objects remain in memory for a longer time span (Figure 3).

Simulators and Production

A key capability of perceptual memory is the production of new perceptual patterns and concepts. Using simulation, the activation of a perceptual symbol can evoke the construction of both a past and a fictional situation. Citing [Barsalou, 1999]: "Productivity in perceptual symbol systems is approximately the symbol formation process run in reverse. During

[3]Such an approach could explain why certain marginal perceptions are retained for a long time if they occurred in a traumatic setting.

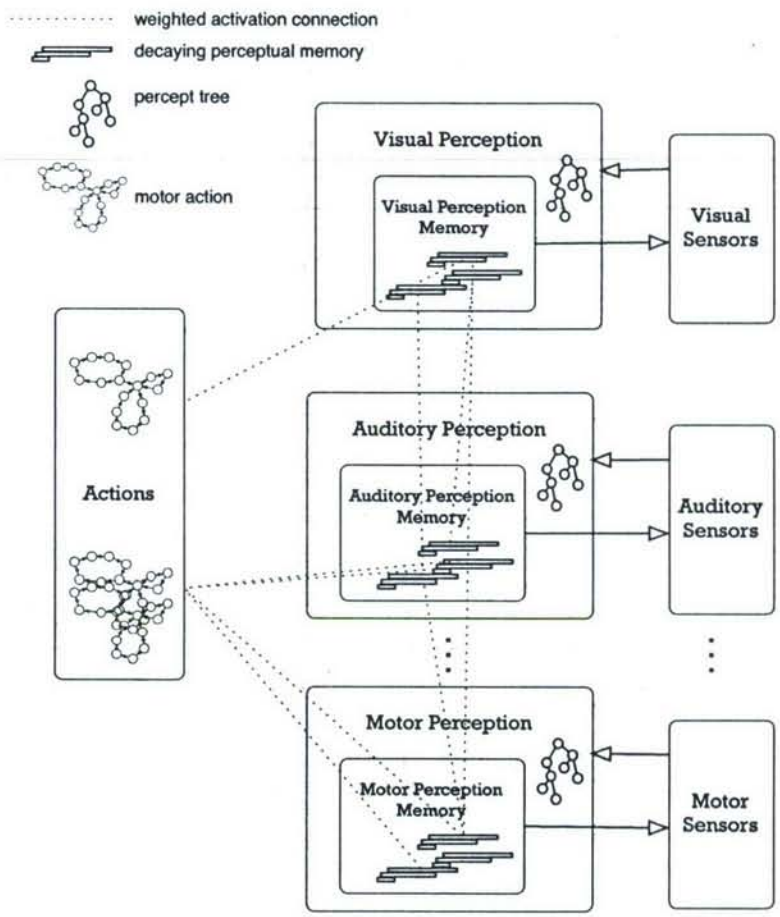

Figure 2: Perception-Based Memory

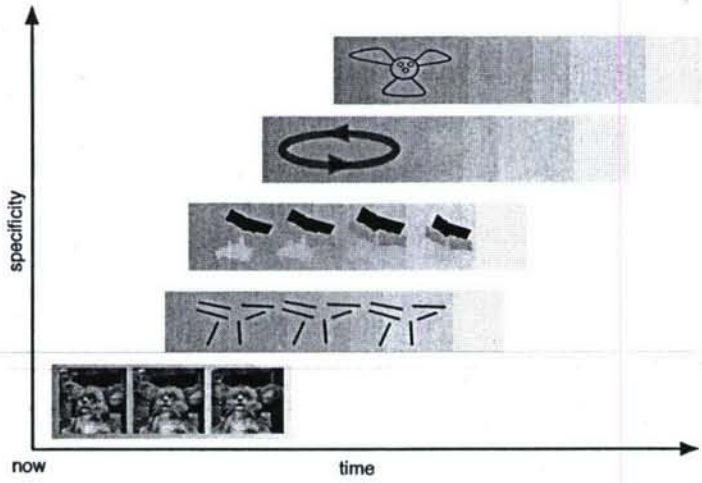

Figure 3: Perceptual memory is filtered up in specificity and remains in memory for a variable amount of time, based—among others—on the information requirements at each specificity level.

symbol formation, large amounts of information are filtered out of perceptual representations to form a schematic representation of a selected aspect. During productivity, small amounts of the information filtered out are added back."

One mechanism enabling the top-down processing of perceptual memory is the implicit connection between feature detecting percepts within a certain modality. As illustrated in Figure 4, an activation of a *red* detector in a perceptual concept implicitly activates other percepts which have activated the same node (as these percept nodes model the same underlying neural structure). Based on the weight of the connection between the activated feature and the perceptual snapshot, different concepts are activated to a varying extent.

A second activation mechanism operates on a higher level, linking percepts of different modalities using weighted connections (Figure 5). Weights of connectivity are learned and refined over time. Their value is influenced by a number of factors, such as frequency of co-ocurrance, attention while creating the perceptual memory, and affective states during perception.

In order to resolve a multitude of (possibly contradictory) perceptual activations using limited processing capacity, perceptual memories can inhibit one another. A more dominant activation of a certain perceptual symbol (for example — the sound "nose" activating imagery of a human nose rather than an airplane's nose) will inhibit a less dominant activation. However, to enable less dominant activations after a certain time (based, for example,

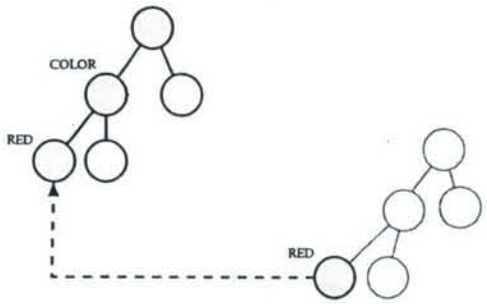

Figure 4: Simulation activates percepts if an active node was triggered in another percept.

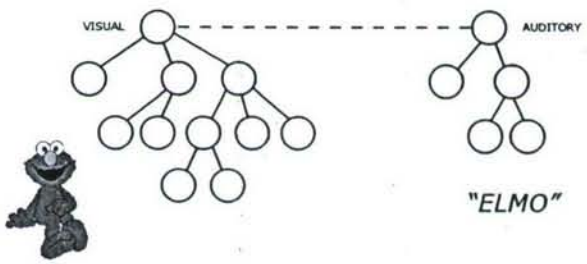

Figure 5: Percept trees in different modalities are interconnected.

on the monitoring of higher level goals), a self-inhibiting process can be used to decay the dominance of active perceptual symbols.

As part of the research architecture proposed here, we have begun implementing a perception-based associative memory model which supports top-down biasing of lower-level perceptual models. In the example depicted in Figure 7, an auditory percept (the sound "Elmo") activates a canonical visual memory of the figure, which — using the same pathway utilized in visual perception — detects the dominant color of the image. This color is then used as a bias affecting the low-level visual buffer, shifting it towards detection of similarly colored areas, eventually priming the system to detect the Elmo puppet in the visual field more easily. This alternative interpretation of the concept of anticipatory behavior illustrates the potential perceptual capabilities of a simulator-based memory model.

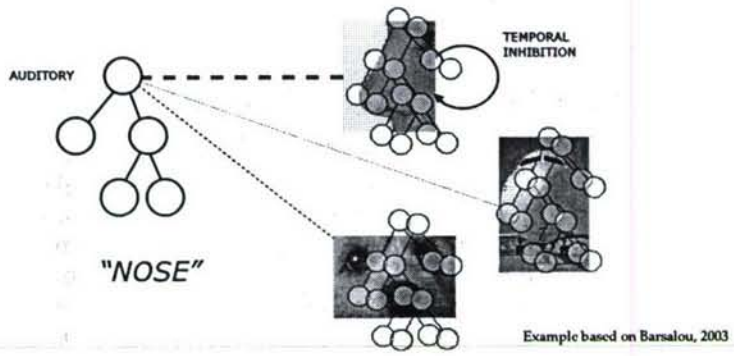

Figure 6: Competing perceptual symbols are incrementally resolved using dominant symbol self-inhibitory decay.

Emulators

Based on behavioral and neurological finding in humans — and of their apparent importance to joint action — perceptual activation and simulation does not only occur in retrospect, but also prospectively, in forms of so-called "emulators" [Wilson and Knoblich, 2005]. In the proposed system, emulators are modeled as a parallel perceptual activation system projecting expectations of perceptual experiences at an atomic time scale. Emulators can also run on representations of procedural knowledge to engage in plan recognition when observing a conspecific, or to anticipate the potential outcome of its own actions.

Action-Perception Activation Networks

Not only perceptions are situated in weighted activation networks, but these are also connected to motor actions (see: Figure 2). This is in line with the action-perception integration approach laid out earlier. Inspired among others by architectures such as Contention Scheduling [Cooper and Shallice, 2000], activities and perceptual concepts occur in a perpetually updating activation relationship.

To elaborate on the operation of such networks: current perceptions exert a weighted influence on activities, leading both to a potential action selection, as well as to the activation of an attention mechanism which in turn refines the perceptual task, and aids its success. Thus, for example, the presentation of a screwdriver may activate a grasping action, as well as the activity of applying a screwdriver to a screw. This will activate a perceptual simulator guiding the visual search towards a screw-like object, and the motor memory related to a clockwise rotation of the wrist.

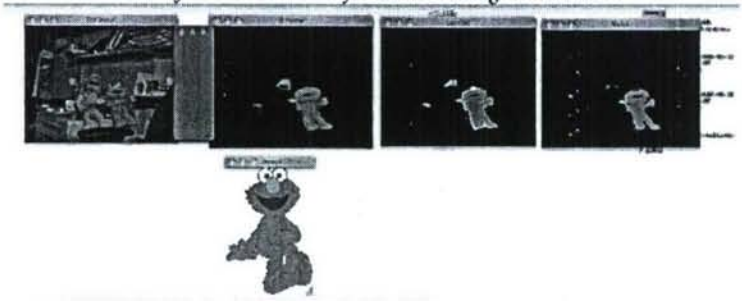

Figure 7: Top-down processing and cross-modal activation in a perceptual-based associative memory model.

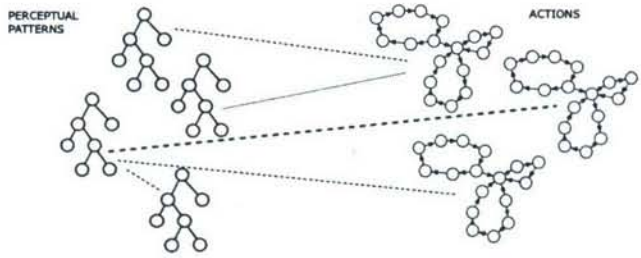

Figure 8: Perceptual symbols and motor activity are interconnected with weighted activations.

In accordance with behavioral evidence [McLeod and Posner, 1984, Wilson, 2001], certain modalities of perception are more strongly connected to particular motor subsystems, resulting in so-called *privileged loops between perception and action*. Disregarding to the discussion of innateness or development of such connections, these "loops" are found in humans to exist, for example, between auditory perception and speech production, as well as between the perception of conspecifics' actions and own body motor system. This thesis proposes that these mechanisms play a significant role in practiced activity and are instrumental to the quick response time necessitated by fluent joint action.

Specifically, it is a hypothesis of this thesis, that such privileged loops as well as other highly weighted connections within and between perceptual and motor activation patterns give rise to some of the nonverbal behavior that is necessary for fluent joint action. The

time-shortest path from a perception to an action might "lead through" a certain set of perceptual and motor activation, which may double as a coordination device for the swift success of a joint activity.

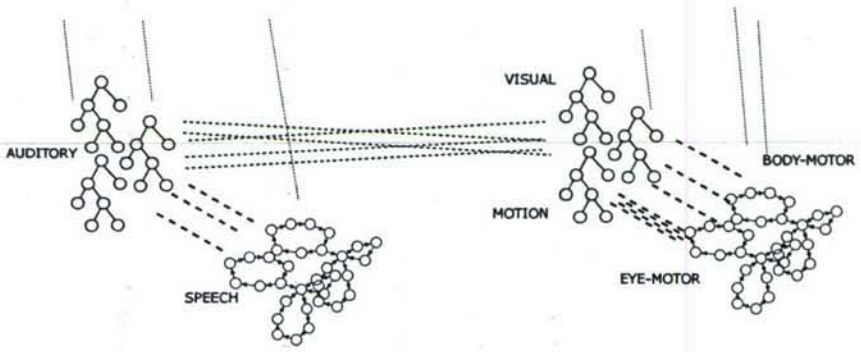

Figure 9: Stronger connections between perception and action result in so-called *privileged loops*.

Prediction and Anticipated Action

Among other factors, successful coordinated action has been linked to the formation of expectations of each partner's actions by the other [Flanagan and Johansson, 2003], and the subsequent acting on these expectations [Knoblich and Jordan, 2003]. We have presented initial work aimed at understanding possible mechanisms to achieve this behavior in a collaborative robot [Hoffman and Breazeal, 2006]. A further research goal is to evaluate whether more fluid anticipatory action can emerge from an approach that uses perception-action activation as its underlying principle, and what directive role anticipation can play in the formation of perceptual procession and fluent joint action.

In the proposed framework, anticipation appears on the perceptual level by the use of simulators, biasing incoming perceptual data; and emulators, predictively building perceptual models of temporal sequences. With the introduction of action sequences, anticipation can also be formed by the cross-activation of perception and action, possibly using a Bayesian model of action transitions, and preemptively activating perceptual symbols based on the anticipation of actions in such sequences.

By modeling goals and intentions of the human collaborator, using a simulation-theoretic framework (as, for example, described in [Gray, 2004]), anticipation can occur on higher task levels, "trickling down" through action detection mechanisms to the action-perception activation network.

Self As Simulator for Mindreading

Simulation Theory (ST) is one of the dominant hypotheses about the nature of the cognitive mechanisms that underlie theory of mind [Davies and Stone, 1995, Gordon, 1986, Heal, 2003]. Simulation Theory posits that by simulating another persons actions and the stimuli they are experiencing using our own behavioral and stimulus processing mechanisms, humans can make predictions about the behaviors and mental states of others based on the mental states and behaviors that we would possess in their situation. In short, by thinking as if we were the other person, we can use our own cognitive, behavioral, and motivational systems to understand what is going on in the heads of others.

Andrew Meltzoff proposes that the way in which infants learn to simulate others is through imitative interactions. Meltzoff posits that infants are in fact intrinsically motivated to imitate their conspecifics, and that the act of successful imitation is its own reward.

For instance, Meltzoff hypothesizes that the human infants ability to translate the perception of anothers action into the production of their own action provides a basis for learning about self-other similarities, and for learning the connection between behaviors and the mental states producing them [Meltzoff, 1996]. To begin with, imitating anothers expression or movement is a literal simulation of their behavior. By physically copying what the adult is doing, the infant must, in a primitive sense, generate many of the same mental phenomena the adult is experiencing, such as the motor plans for the movement. Meltzoff notes that the extent to which a motor plan can be considered a low-level intention, imitation provides the opportunity to begin learning connections between perceived behaviors and the intentions that produce them. Emotional empathy is one of the earliest forms of social understanding that imitation could facilitate. Experiments have shown that producing a facial expression generally associated with a particular emotion is sufficient for eliciting that emotion [Strack et al., 1988]. Hence, simply mimicking the facial expressions of others could cause the infant to feel what the other is feeling, thereby allowing the infant to learn how to interpret emotional states of others from facial expressions and body language.

Mirror neurons have been proposed as a possible neurological mechanism underlying both imitative abilities and Simulation Theory-type prediction of others behaviors and mental states [Gallese and Goldman, 1998]. Within area F5 of the monkeys premotor cortex, these neurons show similar activity both when a primate observes a goal-directed action of another (such as grasping or manipulating an object), and when it carries out that same goal-directed action [Rizzolatti et al., 1996, Gallese et al., 1996]. This firing pattern has led researchers to hypothesize that there exists a common coding between perceived and generated actions. Meltzoff argues that this structure is represented within an intermodal space into which infants are able to map all expressions and movements that they perceive, regardless of their source. In other words, the intermodal space functions as a universal format for representing gestures and poses—those the infant feels himself doing, and those he sees the adult carrying out. The universal format is in terms of the movement primitives within his act space. Thus the perceived expression is translated into the same movement representation that the infants motor system uses (recall the discussion of mirror neurons in

section 3.3) making their comparison much simpler. The imitative link between movement perception and production is forged in the intermodal space.

Inspired by these theories and findings, our simulation-theoretic approach and implementation enables the cognitive agent to monitor an adjacent conspecific (i.e., the human trainer in the simulation) by simulating his or her behavior within the agent's own generative mechanisms on the motor, goal-directed behavior, and perceptual-belief levels. Our implementation computationally models simulation-theoretic mechanisms throughout several systems within the overall cognitive architecture. For instance, within the motor system, mirror-neuron inspired mechanisms are used to map and represent perceived body positions of another into the cognitive agent's own joint space to conduct action recognition. The agent reuses its belief-construction systems from the visual perspective of the human-teacher to predict the beliefs the human is likely to hold to be true given what he or she can visually observe. Finally, within the goal-directed behavior system, where schemas relate preconditions and actions with desired outcomes and are organized to represent hierarchical tasks, motor information is used along with perceptual and other contextual clues (i.e., task knowledge) to infer the human's goals and how he or she might be trying to achieve them (i.e., plan recognition).

The general methodology is summarized as follows:

- Map the human's actions onto the cognitive agent's motor representations via the mirror system within the cognitive agent's perception-activation networks. Tag these actions as coming from "other."
- Use dual-pathways from the motor to the cognitive systems to pass this information to those systems that evoke these movements. For instance, movements used to achieve task goals would pass up to the goal-directed behavior system that run emulators over procedural representations.
- Consider the perceptual context from the human's visual perspective. Based on this, hypothesize what their likely beliefs are about the perceptual context by running simulators in the perceptual activation networks. Tag these as coming from "other".
- Consider any other relevant contextual information, such as task knowledge.
- With this dual-pathway information, use the cognitive systems as a emulator running on these "other-derived" inputs to infer the likely goals or beliefs that would arise given these circumstances within the associated systems. (Note that multiple hypotheses could be generated and weighted probabilistically to indicate how likely they are).
- Use this information to predict the human's behavior, and to shape the agent's own responses.
- Incorrect inferences present an opportunity for learning. Ideally, even incorrect inferences should at some level seem plausible to the human. This will assist with efficient error recovery and reduce the chances the same mistake is made in the future.

Intentions, Motivations, and Supervision

It should be noted that much of the above architecture is predominantly apt for governing automatic or routine activity. Perceptions precipitate concepts and object memories, which in turn prompt actions governing future perception and motor activation. While this model can be adequate for well-established behavior, an agent acting jointly with a human must also behave in concordance with internal motivations and intentions, and higher-level supervision. This is particularly crucial because humans naturally assign internal states and intentions to animate and even inanimate objects [Baldwin and Baird, 2001, Dennett, 1987, Malle et al., 2001]. Our cognitive agent acting with people must behave according to internal drives as well as clearly communicate these drives to their human counterpart.

In the proposed architecture, supervisory intention- and motivation-based or affect-based control affects the autonomic processing scheme outlined above (Figure 10). At the base of this supervisory system lie core drives, such as hunger, boredom, attention-seeking, and other domain-specific drives (such as task success), modeled as scalar fluents, which the agent seeks to maintain at an optimal level (similar to [Breazeal, 2002]). Emotional factors can also contribute at this level. If any of those fluents falls above or below the defined range, they trigger a goal request. The number and identity of the agent's motivational drives are fixed, but the emotional system adds flexibility and adaptabilitiy. The Goal Arbitration module decides what goal(s) to pursue based on situational context and internal state (e.g., current goals, motivations and emotions).

Once a goal is selected, a deliberative processes follow to construct a plan that satisfies the goal. A plan can then activate or override the automatic action selection mechanisms in the action-perception activation network, activating simulators and guiding perception and motor activity, according to the principles laid out in [Cooper and Shallice, 2000]. The outcome of actions is fed back through perceptual acquisition to the motivational system. Note that this model can also be used as a basis for an experienced-based intention reading framework, as outlined in the previous section

Practice

According to the model put forth here, practice operates on two levels: it alters the weights of connections within the perception and motor activation networks and thus transfers deliberate action selection procedures (stemming from the motivational layer) to a faster action activation mechanism; practice also helps shape the anticipatory action selection mechanism by reinforcing transition probabilities and time estimates which in turn are being used by emulators to predict future perceptual and motor operations.

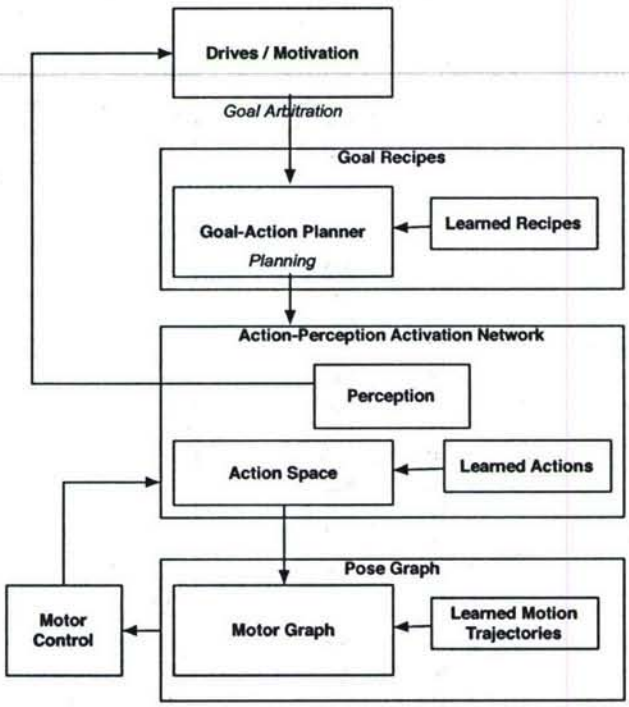

Figure 10: Motivational drives are used to form plans overriding or swaying action selection in the Action-Perception Activation Network.

4 Conclusion

As part of our phase I effort, we began to computationally implement this novel biologically inspired cognitive architecture described in this document. It is the first attempt to computationally model these Barsalou-like simulators and test them within behaving cognitive agents, namely robots.

We have tested an early version of this architecture on two physical robot systems: Leonardo (a 65 degree of freedom humanoid robot), and AUR (a 6 degree of freedom robotic desk lamp). Both have been used to explore the dynamics of this architecture as it applies to improving the fluency of joint action through practice. In the case of Leonardo, a game of patty-cake was used to test the robot's use of emulators to predict and anticipate a sequence of actions practiced with a human. In the case of AUR, the same architecture was used to have the robotic lamp anticipate what part of a dimly lit workspace to illuminate based on the human's needs during a collaborative task. The final results of this effort shall be published as part of Guy Hoffman's PhD Thesis: *Fluency and Embodiment for Robots Acting with Humans* to be completed in August 2007.

References

[Baldwin and Baird, 2001] Baldwin, D. and Baird, J. (2001). Discerning intentions in dynamic human action. *Trends in Cognitive Sciences*, 5(4):171–178.

[Barsalou, 1999] Barsalou, L. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22:577–660.

[Barsalou et al., 1999] Barsalou, L., Solomon, K., and Wu, L. (1999). Perceptual simulation in conceptual tasks. In Hiraga, M. K., Sinha, C., and Wilcox, S., editors, *Cultural, Typological, and Psychological Perspectives in Cognitive Linguistics: The Proceedings of the 4th Conference of the International Cognitive Linguistics Association*, volume 3, Amsterdam.

[Blumberg et al., 2002] Blumberg, B., Downie, M., Ivanov, Y., Berlin, M., Johnson, M., and Tomlinson, B. (2002). Integrated learning for interactive synthetic characters. In *Proceedings of the ACM SIGGRAPH*.

[Breazeal, 2002] Breazeal, C. (2002). *Designing Sociable Robots*. MIT Press.

[Bregler, 1997] Bregler, C. (1997). Learning and recognizing human dynamics in video sequences. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*, page 568. IEEE Computer Society.

[Chovil, 1992] Chovil, N. (1992). Discourse-oriented facial displays in conversation. *Research on Language and Social Interaction*, 25:163–194.

[Cooper and Shallice, 2000] Cooper, R. and Shallice, T. (2000). Contention scheduling and the control of routing activities. *Cognitive Neuropsychology*, 17:297–338.

[Davies and Stone, 1995] Davies, M. and Stone, T. (1995). Introduction. In Davies, M. and Stone, T., editors, *Folk Psychology: The Theory of Mind Debate*. Blackwell, Cambridge.

[Dennett, 1987] Dennett, D. C. (1987). Three kinds of intentional psychology. In *The Intentional Stance*, chapter 3. MIT Press, Cambridge, MA.

[Flanagan and Johansson, 2003] Flanagan, J. R. and Johansson, R. S. (2003). Action plans used in action observation. *Nature*, 424(6950):769–771.

[Gallese et al., 1996] Gallese, V., Fadiga, L., Fogassi, L., and Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119:593–609.

[Gallese and Goldman, 1996] Gallese, V. and Goldman, A. (1996). Mirror neurons and the simulation theory of mind-reading. *Brain*, 2(12):493–501.

[Gallese and Goldman, 1998] Gallese, V. and Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, 2(12):493–501.

[Gordon, 1986] Gordon, R. (1986). Folk psychology as simulation. *Mind and Language*, 1:158–171.

[Gray, 2004] Gray, J. (2004). Goal and action inference for helpful robots using self as simulator. Master's thesis, Massachusetts Institute of Technology.

[Hamdan et al., 1999] Hamdan, R., Heitz, F., and Thoraval, L. (1999). Gesture localization and recognition using probabilistic visual learning. In *Proceedings of the 1999 Conference on Computer Vision and Pattern Recognition (CVPR '99)*, pages 2098–2103, Ft. Collins, CO, USA.

[Heal, 2003] Heal, J. (2003). Understanding other minds from the inside. In *Mind, Reason and Imagination*, pages 28–44. Cambridge University Press, Cambridge UK.

[Hoffman and Breazeal, 2006] Hoffman, G. and Breazeal, C. (2006). Cost-based anticipatory action-selection for human-robot fluency. Unpublished.

[Knoblich and Jordan, 2003] Knoblich, G. and Jordan, J. S. (2003). Action coordination in groups and individuals: learning anticipatory control. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(5):1006–1016.

[Kosslyn, 1995] Kosslyn, S. (1995). Mental imagery. In Kosslyn, S. and D.N.Osherson, editors, *Invitation to Cognitive Science: Visual Cognition*, volume 2, chapter 7, pages 276–296. MIT Press, Cambridge, MA, 2nd edition.

[Krauss et al., 1996] Krauss, R. M., Chen, Y., and Chawla, P. (1996). Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us? In Zanna, M., editor, *Advances in experimental social psychology*, pages 389–450. Tampa: Academic Press.

[Locke and Kutz, 1975] Locke, J. L. and Kutz, K. J. (1975). Memory for speech and speech for memory. *Journal of Speech and Hearing Research*, 18:176–191.

[Malle et al., 2001] Malle, B., Moses, L., and Baldwin, D., editors (2001). *Intentions and Intentionality*. MIT Press.

[Martin, 2001] Martin, A. (2001). Functional neuroimaging of semantic memory. In Cabeza, R. and A. Kingstone Kingstone, A., editors, *Handbook of Functional Neuroimaging of Cognition*, pages 153–186. MIT Press.

[Massaro and Cohen, 1983] Massaro, D. and Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 9(5):753–771.

[Matarić, 2002] Matarić, M. J. (2002). Sensory-motor primitives as a basis for imitation: linking perception to action and biology to robotics. In Dautenhahn, K. and Nehaniv, C. L., editors, *Imitation in animals and artifacts*, chapter 15, pages 391–422. MIT Press.

[McClave, 2000] McClave, E. (2000). Linguistic functions of head movements in the context of speech. *Journal of Pragmatics*, 32:855–878.

[McLeod and Posner, 1984] McLeod, P. and Posner, M. I. (1984). Privileged loops from percept to act. In Bouma, H. and Bouwhuis, D. G., editors, *Attention and performance*, volume 10, pages 55–66. Erlbaum, Hillsdale, NJ.

[Meltzoff, 1996] Meltzoff, A. N. (1996). The human infant as imitative generalist: A 20-year progress report on infant imitation with implications for comparative psychology. In C. M. Heyes, B. G., editor, *Social Learning in Animals: The Roots of Culture*. Academic Press, San Diego, CA.

[Meltzoff and Moore, 1997] Meltzoff, A. N. and Moore, M. K. (1997). Explaining facial imitation: a theoretical model. *Early Development and Parenting*, 6:179–192.

[Rich et al., 2001] Rich, C., Sidner, C. L., and Lesh, N. (2001). Collagen: Applying collaborative discourse theory to human-computer collaboration. *AI Magazine*, 22(4):15–25.

[Rizzolatti et al., 1996] Rizzolatti, G., Fadiga, L., Gallese, V., and Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141.

[Sebanz et al., 2006] Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2):70–76.

[Somerville et al., 2004] Somerville, J. A., Woodward, A. L., and Needham, A. (2004). Action experience alters 3-month-old infants' perception of others actions. *Cognition*.

[Spivey et al., 2005] Spivey, M. J., Richardson, D. C., and Gonzalez-Marquez, M. (2005). On the perceptual-motor and image-schematic infrastructure of language. In Pecher, D. and Zwaan, R. A., editors, *Grounding cognition: the role of perception and action in memory, language, and thinking*. Cambridge Univ. Press, Cambridge, UK.

[Stanfield and Zwaan, 2001] Stanfield, R. and Zwaan, R. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, 12:153–156.

[Strack et al., 1988] Strack, F., Martin, L., and Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology*, 54:768777.

[Thornton et al., 1998] Thornton, I., J., P., and Shiffrar, M. (1998). The visual perception of human locomotion. *Cognitive Neuropsychology*, 15:535–552.

[Wilson, 2001] Wilson, M. (2001). Perceiving imitable stimuli: consequences of isomorphism between input and output. *Psychological Bulletin*, 127(4):543–553.

[Wilson and Knoblich, 2005] Wilson, M. and Knoblich, G. (2005). The case for motor involvement in perceiving conspecifics. *Psychological Bulletin*, 131:460–473.

[Wilson et al., 2004] Wilson, S., Saygin, A., Sereno, M., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nature Neuroscience*, 7:701–702.