A Cognitive Architecture that Combines Internal Simulation with a Global Workspace

Murray Shanahan

Dept. Electrical & Electronic Engineering,
Imperial College,
Exhibition Road,
London SW7 2BT,
England.
m.shanahan@imperial.ac.uk

Abstract

This paper proposes a brain-inspired cognitive architecture that incorporates approximations to the concepts of consciousness, imagination, and emotion. To emulate the empirically established cognitive efficacy of conscious as opposed to non-conscious information processing in the mammalian brain, the architecture adopts a model of information flow from global workspace theory. Cognitive functions such as anticipation and planning are realised through internal simulation of interaction with the environment. Action selection, in both actual and internally simulated interaction with the environment, is mediated by affect. An implementation of the architecture is described which is based on weightless neurons and is used to control a simulated robot.

1 Introduction

Cotterill (1998; 2001) advances the proposal that thought is "internally simulated interaction with the environment", and Hesslow (2002) argues that this "simulation hypothesis" can explain our experience of an inner world. However, while the simulation hypothesis has the potential to account for the content of conscious thought, it doesn't supply an answer to the question of what it is that distinguishes conscious from non-conscious activity in the brain. By contrast, global workspace theory can account for this distinction by appealing to an information processing architecture that features both competition among, and broadcast to, different brain processes (Baars, 1988; 1997).

The present article effects a marriage between these two proposals by presenting a neural-level cognitive architecture that realises an internal sensorimotor loop in which information passes through multiple competing cortical areas and a global workspace. This architecture, whose implementation is described here along with its application to the control of a simulated robot, serves to demonstrate that i) the simulation hypothesis can be elegantly reconciled with global workspace theory, and ii) a robot controller which draws on these contemporary ideas from the scientific study of consciousness is also viable from an engineering point of view.

From its inception to the present day, mainstream cognitive science has assumed language and reason to be the right conceptual foundations on which to build a scientific understanding of cognition. By contrast, the champions of biologically-inspired AI jetisoned these concepts in the 1990s. But at the same time they abandoned the very idea of cognition as a primary object of study. The present paper takes it for granted that understanding cognition will be central to achieving human-level artificial intelligence. However, the brain-inspired architecture described here, instead of manipulating declarative, language-like representations in the manner of classical AI, realises cognitive function through topographically organised maps of neurons, which can be thought of as a form of *analogical* (or *diagrammatic* or *iconic*) representation whose structure is close to that of the sensory input of the robot whose actions they mediate (Sloman, 1971; Glasgow, *et al.*, 1995).

Analogical representations are especially advantageous in the context of spatial cognition, which, though not the focus of the present paper, is a crucial capacity for both animals and robots. While common sense inferences about shape and space are notoriously difficult with traditional logic-based approaches (Shanahan, 2004), in an analogical representation basic spatial properties such as distance, size, shape, and location are inherent in the medium itself and require negligible computation to extract. Furthermore, traditional language-like representations bear a subtle and contentious relationship to the world they are supposed to represent, and raise difficult questions about intentionality and symbol grounding (Harnad, 1990; Shanahan, 2005). With analogical representations, which closely resemble raw sensory input, this semantic gap is small and these questions are more easily answered.

In addition to these representational considerations, the design of the proposed architecture reflects the view, common among proponents of connectionism, that parallel computation should be embraced as a foundational concept rather than sidelined as a mere implementation issue. Specifically, the present paper advocates a computational architecture based on the *global workspace* model of information flow, in which a serial procession of states emerges from the interaction of many separate, parallel processes (Baars, 1988; 2002). This *serial* procession of states, which includes the unfolding of conscious content in human working memory (Baars, & Franklin, 2003), facilitates anticipation and planning and enables a cognitively-enhanced form of action selection. Yet the robustness and flexibility of these cognitive functions depends on the behind-the-scenes performance of extremely large numbers of *parallel* computations, only the most relevant of which end up making a contribution to the ongoing serial thread (Shanahan & Baars, 2005).

The architecture presented here is intended to be neurologically plausible at the level of large-scale neural assemblies, and contains analogues of a variety of brain structures and systems, including multiple motor-cortical populations (that compete for access to the global workspace), internal sensorimotor loops (capable of rehearsing trajectories through sensorimotor space), the basal ganglia (to carry out action selection), and the amygdala (to guide action selection through affect). But the central component is the global workspace itself, for which there are a number of candidate homologues in the vertebrate brain, including higher-order thalamocortical relays, and long-range corticocortical fibres.

In its overall conception, the architecture appeals to the notions of imagination and emotion as well as consciousness. Although perhaps only rough approximations to their humanly-applicable counterparts, the way these concepts are deployed here is inspired by their increasingly important role in the brain sciences (Damasio, 2000). As such, the architecture described builds on the work of a number of other authors who have applied these ideas in the context of robotics or artificial intelligence.

- **Consciousness** As already touched on, global workspace theory proposes a model of information flow in which conscious information processing is cognitively efficacious because it integrates the results of the brain's massively parallel computational resources (Baars, 1988; 2002). The global workspace architecture has previously been used in the design of software

agents (Franklin & Graesser, 1999; Franklin, 2003), but its application to robotics has so far been neglected in the literature.

- **Imagination** Although far from mainstrem, the view that thought is internally simulated interaction with the environment or, to put it another way, the rehearsal of trajectories through sensorimotor space prior to their possible enactment, has influenced a number of robotics researchers in the biologically-inspired tradition, including Chrisley (1990), Stein (1995), Holland (2003), Hoffmann & Möller (2004), and Ziemke, *et al*. (2005).
- **Emotion** Based on clinical studies, Damasio (1995) argued persuasively that the human capacity for rational deliberation is dependent on an intact affective system, and many other cognitive scientists subscribe to the view that affect addresses the problems of decision making and action selection (Picard, 1997; Sloman, 2001). It permits a number of factors to be blended together and brought to bear on the problem of contention for resources (ie: muscles) by different brain processes. Neurologically plausible mechanisms of action selection compatible with this idea have been demonstrated in a robotics setting by Prescott, *et al* (1999) and Cañamero (2003).

The rest of the paper is organised as follows. Section 2 presents a top-level schematic of the architecture, distinguishing first-order (external) from higher-order (internal) sensorimotor loops, and goes on to describe the roles of affect and the global workspace in the architecture. Section 3 describes the implementation of the architecture at the level of its various neural assemblies and the circuitry that interconnects them. Section 4 outlines some experimental results obtained with the system, both with and without external connection to the robot. The concluding discussion in section 5 addresses some of the methodological and philosophical issues the work throws up.

2 A Top-level Schematic

Fig. 1 shows a top-level schematic of the architecture. It can be thought of in terms of two interacting sub-systems. The first-order system is purely reactive, and determines an immediate motor response to the present situation without the intervention of cognition. But these unmediated motor responses are subject to a veto imposed by BG (the basal ganglia analogue). Through BG, which carries out salience-based action selection, the higher-order loop modulates the behaviour of the first-order system. It does this by adjusting the salience of currently executable

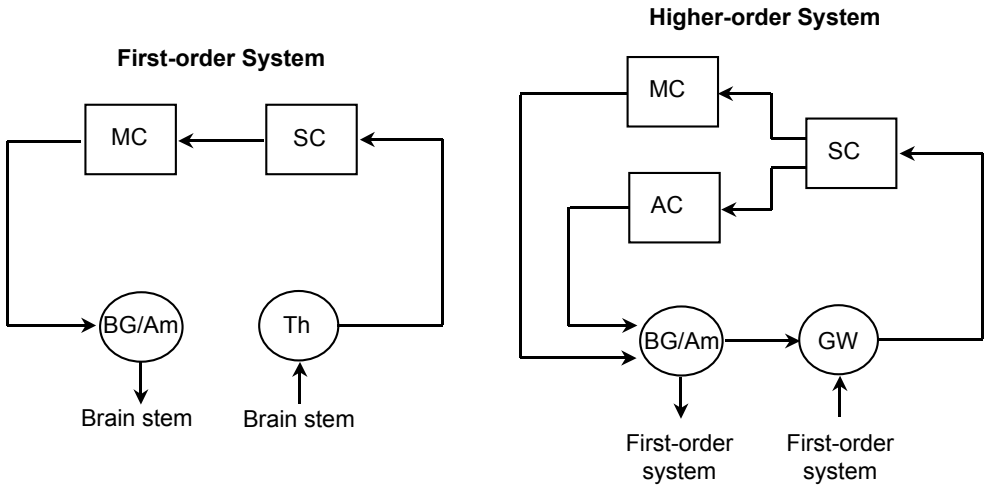

Fig. 1: A top-level schematic of the architecture. MC = motor cortex, SC = sensory cortex, AC = association cortex, BG = basal ganglia, Am = amygdala, Th = thalamus, GW = global workspace.

actions. Sometimes this adjustment will have no effect. But sometimes it will result in a new action becoming the most salient. And sometimes it will boost an action's salience above the threshold required to release its veto, bringing about that action's execution.

The higher-order system computes these salience adjustments by carrying out off-line rehearsals of trajectories through (abstractions of) the robot's sensorimotor space. In this way – through the exercise of its "imagination" – the robot is able to anticipate and plan for potential rewards and threats without exhibiting overt behaviour.

The first- and higher-order systems have the same basic components and structure. Both are sensorimotor loops. The key difference is that the first-order loop is closed through interaction with the world itself while the higher-order loop is closed internally. This internal closure is facilitated by AC, which simulates — or generates an abstraction of — the sensory stimulus expected to follow from a given motor output, and fulfils a similar role to a *forward model* in the work of various authors (Demiris & Hayes, 2002; Wolpert, *et al.*, 2003; Grush, 2004). The cortical components of the higher-order system (SC, AC, and MC) correspond neurologically to regions of association cortex, including the prefrontal cortex which is implicated in planning and working memory (Fuster, 1997).

2.1 Affect and Action Selection

Analogues of various sub-cortical and limbic structures appear in both the first- and higher-order systems, namely the basal ganglia, the amygdala, and the thalamus. In both systems, the basal ganglia are involved in action selection. Although, for ease of presentation, the schematic in Fig. 1 suggests that the final stage of motor output before the brain stem is the basal ganglia, the truth is more complicated in both the mammalian brain and the robot architecture it has inspired.

In the mammalian brain, the pertinent class of basal ganglia circuits originate in cortex, then traverse a number of nuclei of the basal ganglia, and finally pass through the thalamus on their way back to the cortical site from which they originated. The projections up to cortex are thought to effect action selection by suppressing all motor output except for that having the highest salience, which thereby makes it directly to the brain stem and causes muscular movement (Mink, 1996; Redgrave, *et al.*, 1999). The basolateral nuclei of the amygdala are believed to modulate the affect-based salience information used by the basal ganglia through the association of cortically mediated stimuli with threat or reward (Baxter & Murray, 2002; Cardinal, *et al.*, 2002).

The robot architecture includes analogues of the basal ganglia and amygdala that function in a similar way. These operate in both the first- and higher-order systems. In the first-order system, the amygdala analogue associates patterns of thalamocortical activation with either reward or punishment, and thereby modulates the salience attached to each currently executable action. The basal ganglia analogue adjudicates the competition between each executable action and, using a winner-takes-all strategy, selects the most salient for possible execution. While the salience of the selected action falls below a given threshold it is held on veto, but as soon as its salience exceeds that threshold it is executed.

The roles of the basal ganglia and amygdala analogues in the higher-order system are similar, but not identical, to their roles in the first-order system (Cotterill, 2001). These structures are again responsible for action selection. However, action selection in the higher-order system does not determine overt behaviour but rather selects one path through the robot's sensorimotor space for inner rehearsal in preference to all others. Moreover, as well as gating the output of motor association cortex (MC), the basal ganglia analogue must gate the output

of sensory association cortex (AC) accordingly, and thus determine the next hypothetical sensory state to be processed by the higher-order loop.

This distinction between first-order and higher-order functions within the basal ganglia is reflected in the relevant neuroanatomy. Distinct parallel circuits operate at each level (Nolte, 2002, p. 271). In the first-order circuit, sensorimotor cortex projects to the putamen (a basal ganglia input nucleus), and then to the globus pallidus (a basal ganglia output nucleus), which projects to the ventral lateral and ventral anterior nuclei of the thalamus, which in turn project back to sensorimotor cortex. In the higher-order circuit, association cortex projects to the caudate nucleus (a basal ganglia input structure), and then to the substantia nigra (a basal ganglia output nucleus), which projects to the mediodorsal nucleus of the thalamus, which in turn projects back to association cortex.

2.2 Global Workspace Theory

Global workspace theory advances a model of information flow in which multiple, parallel, specialist processes compete and co-operate for access to a global workspace (Baars, 1988). Gaining access to the global workspace allows a winning coalition of processes to broadcast information back out to the entire set of specialists (Fig. 2). Although the global workspace exhibits a serial procession of broadcast states, each successive state itself is the integrated product of parallel processing.

According to global workspace theory, the mammalian brain instantiates this model of information flow, which permits a distinction to be drawn between conscious and non-conscious information processing. Information that is broadcast via the global workspace is consciously processed while information processing that is confined to the specialists is non-conscious. A considerable body of empirical evidence in favour of this distinction has accumulated in recent years (Baars, 2002).

The particular blend of serial and parallel computation favoured by global workspace theory suggests a way to address the frame problem – in the philosopher's sense of that term (Fodor, 2000) – which in turn suggests that conscious information processing may be cognitively efficacious in a way that non-conscious information processing is not (Shanahan & Baars, 2005). In particular, in the context of so-called informationally unencapsulated cognitive processes, it allows relevant information to be sifted from the irrelevant without

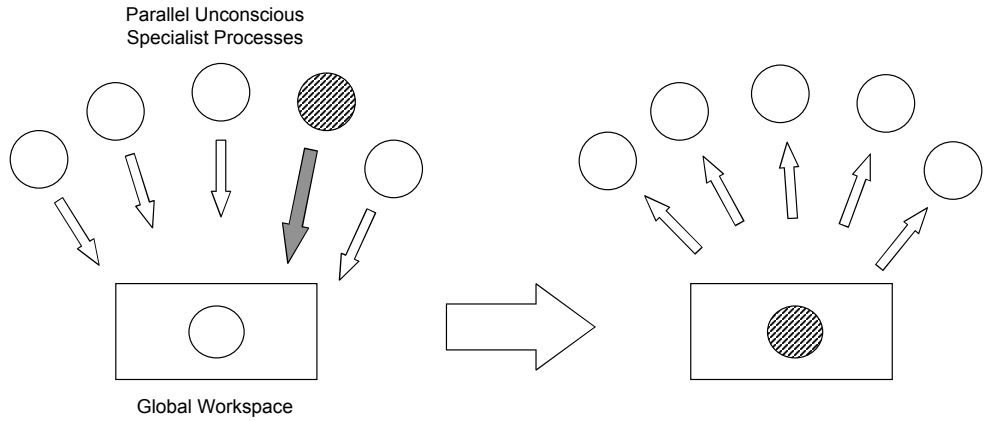

Fig. 2: The global workspace architecture.

incurring an impossible computational burden. More generally, broadcast interleaved with selection facilitates the integration of the activities of large numbers of specialist processes working separately. So the global workspace model can be thought of as one way to manage the massively parallel computational resources that surely underpin human cognitive prowess.

The architecture of this paper conforms to the global workspace model of information flow by incorporating complementary mechanisms for the broadcast of information to multiple cortical areas and for selection between competing patterns of activation within those areas (Fig. 3). As Fig. 3 shows, information fans out from the the global workspace (GW) to multiple cortical sites (within which it may be subject to further local distribution). Conversely, information funnels back into the global workspace, possibly after competition within cortically localised regions (Koch, 2004, ch. 2), thanks to a process of selection between cortical sites realised by the basal ganglia.

In neurological terms, a global workspace might be realised by ascending thalamocortical fibres, or by long-range corticocortical fibres, or by some combination of both. The thalamocortical workspace hypothesis is supported by the fact that the first-order / higher-order distinction is preserved in the thalamus, which contains not only first-order relays that direct signals from the brain stem up to cortex (located, for example, in the lateral geniculate nucleus), but also higher-order relays that route cortical traffic back up to cortex (located, for example, in the pulvinar) (Sherman & Guillery, 2001; 2002).

However, while Sherman and Guillery have suggested a dominant role for thalamocortical relays in cortical communication, others have challenged this

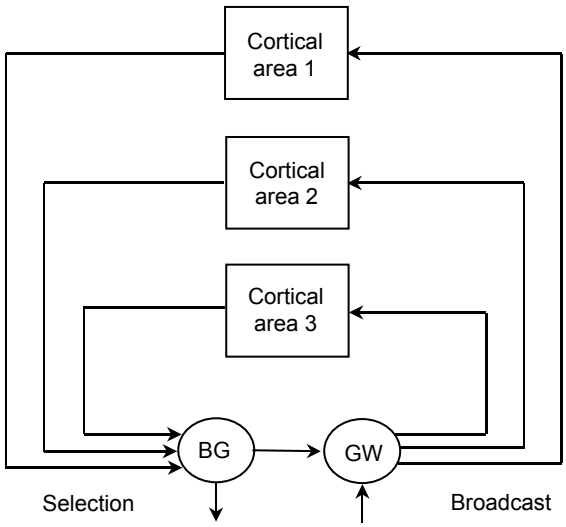

Fig 3: The fan-and-funnel model

position. According to the opposing view, direct corticocortical projections are the main bearers of information from one cortical region to another (Van Essen, 2005). In line with this, Dehaene and his colleagues, who have also implemented a neural-level computer simulation of the global workspace architecture, favour the corticocortical workspace hypothesis (Dehaene, *et al.*, 2003; Dehaene & Changeaux, 2005).

A third possibility is that both hypotheses are true. The tendency for complex biological systems such as the brain to exhibit degeneracy (Edelman & Gally, 2001) suggests that thalamocortical and corticocortical connections might independently be capable of supporting a global workspace. Indeed, the replication principle of Shipp (2003), according to which the connectivity of indirect transthalamic cortical pathways mimics that of direct corticocortical pathways, is supportive of this possibility. However, the present architecture is neutral with respect to this issue, as it does not attempt to model the brain at a sufficiently low-level of detail for the distinction between the thalamocortical and corticocortical hypotheses to make sense.

The fan-and-funnel model of broadcast / distribution and competition / selection depicted in Fig. 3 can be straightforwardly combined with the top-level schematic of Fig. 1, as is apparent from the diagrams. Indeed, the role of the BG component of the higher-order loop introduced in Fig. 1 is precisely to effect a selection between the outputs of multiple competing cortical areas, as shown in Fig. 3. The behaviour of the resulting system is best understood in dynamical

system terms, and in this sense the outlook of the present paper accords with the methodology advocated by Van Gelder (1997).

The global workspace and the competing cortical assemblies each define an attractor landscape. Perceptual categories become attractors in a state space whose structure mirrors that of raw sensory input. Predictions are made by advancing the higher-order sensorimotor system along a simulated trajectory through (an abstraction of) that state space, enabling the global workspace to visit a sequence of attractors. But the various cortical assemblies are coupled in such a way that the vector of motion through this state space is determined by associations between one attractor and another, and different associations compete (and sometimes combine) to influence this vector of motion.

3 An Implementation

The brain-inspired architecture of the previous section has been implemented using NRM (Dunmall, 2000), a tool for building large-scale neural network models using G-RAMs (generalising random access memories) (Figs. 4 and 5). These are weightless neurons employing single-shot training whose update function can be rapidly computed (Aleksander, 1990).

The basic operation of a single G-RAM is illustrated in Fig. 4. The input vector is used to index a lookup table. In the example shown, the input vector of 1011 matches exactly with the fourth line of the table, which yields the output 6. When there is no exact match, the output is given by the line of the lookup table with the smallest Hamming distance from the input vector, so long as this exceeds a predefined threshold. In this example, if the input vector had been 1010, then none of the lines in the lookup table would yield an exact match. But the fourth line would again be the best match, with a Hamming distance of 1, so the output would again be 6. If no line of the lookup table yields a sufficiently close match to the input vector the neuron outputs 0, which represents quiescence.

The implemented system exploits the fact that G-RAMs can be easily organised into attractor networks with similar properties to Hopfield nets (Lockwood & Aleksander, 2003). The core of the implementation, which comprises almost 40,000 neurons and over 3,000,000 connections, is a set of cascaded attractor networks corresponding to each of the components identified in the architectural blueprint of the previous section.

The NRM model is interfaced to Webots, a commercial robot simulation environment (Michel, 2004). The simulated robot is a Khepera with a 64 × 64 pixel camera, and the simulated world contains cylindrical objects of various colours. The Khepera is programmed with a small suite of low-level actions including "rotate until an object is in the centre of the visual field" and "approach

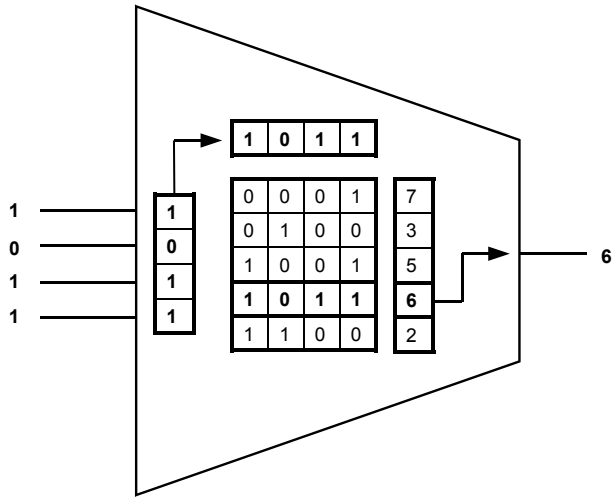

Fig 4: The G-RAM weightless neuron

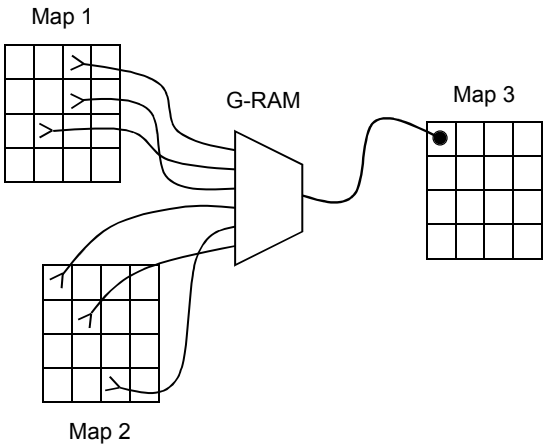

Fig 5: G-RAM maps and connections

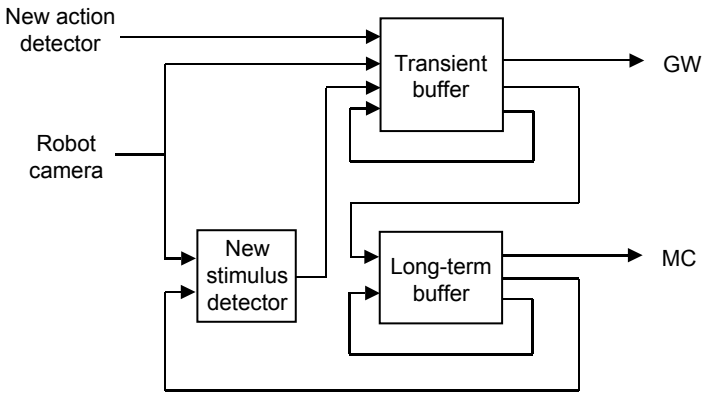

Fig. 6: Visual system circuitry (VC / IT). VC = visual cortex, IT = inferotemporal cortex.

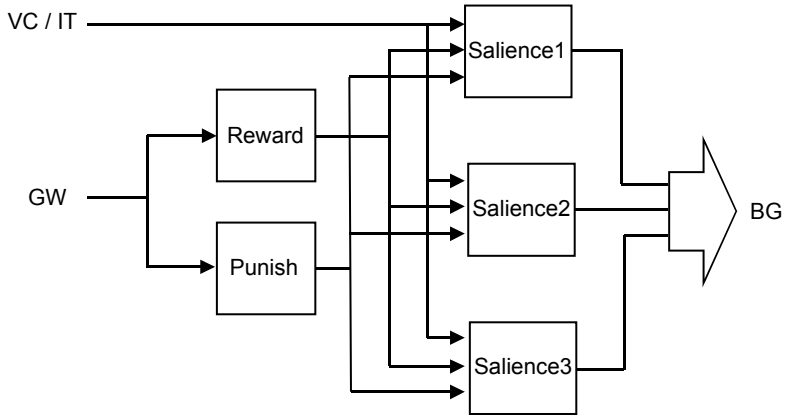

Fig. 7: Affect circuitry (Am)

an object in the centre of the visual field". These two actions alone are sufficient to permit navigation in the robot's simple environment.

The overall system can be divided into four separate modules – the visual system (Fig. 6), the affective system (Fig. 7), the action selection system (Fig. 8), and the broadcast / inner rehearsal system (Fig. 9). Each box in these figures denotes a layer of neurons and each path denotes a bundle of connections. If a path connects a layer A to an $n \times n$ layer B then it comprises n^2 separate pathways – one for each of the neurons in B – each of which itself consists of m input connections originating in a randomly assigned subset of the neurons in A (Fig. 5). For the majority of visual maps m is set to 32.

The two buffers in the visual system comprise 64×64 topographically organised neurons (Fig. 6). These are both attractor networks, a property indicated by the presence of a local feedback path. The transient buffer is activated by the

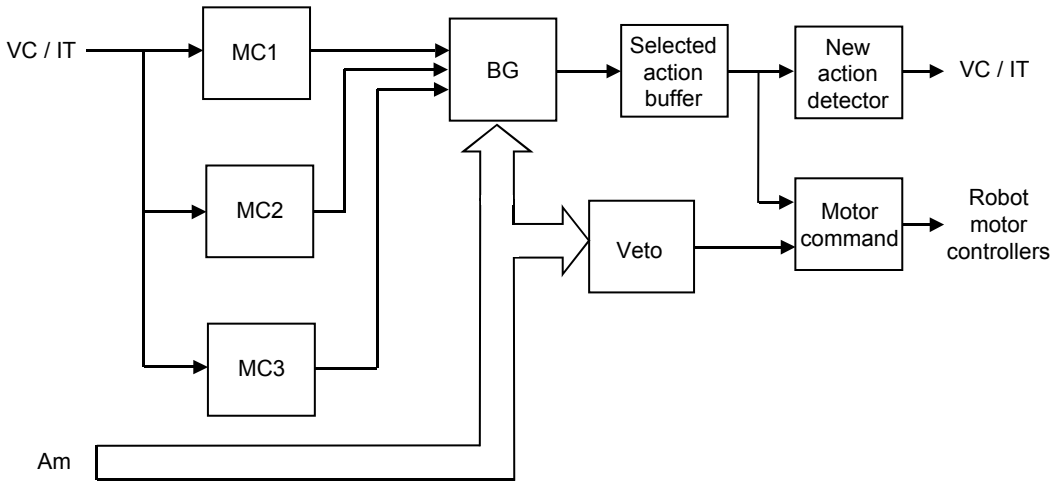

Fig. 8: Action selection circuitry (BG / MC)

presence of a new visual stimulus. The hallmark of a new stimulus is that it can jog the long-term visual buffer out of one attractor and into another. The global workspace component of the inner rehearsal system is loaded from the transient visual buffer, whose contents rapidly fade allowing the dynamics of the higher-order sensorimotor loop to be temporarily dominated by intrinsic activity rather than sensory input.

The contents of the long-term visual buffer are fed to three competing motor-cortical areas, MC1 to MC3 (Fig. 8), each of which responds either with inactivity or with a recommended motor response to the current stimulus. Each recommended response has an associated salience (Fig. 7). This is used by the action selection system to determine the currently most salient action, which is loaded into the "selected action buffer" (Fig. 8). But the currently seleced action is subject to a veto. Only if its salience is sufficiently high does it get loaded into the "motor command" buffer, whose contents is forwarded to the robot's motor controllers for immediate execution.

So far the mechanism described is little different from a standard behaviour-based robot control architecture (Brooks, 1986). What sets it apart from a purely reactive system is its capacity for inner rehearsal. This is realised by the core circuit depicted in Fig. 9. When a new visual stimulus arrives, it overwrites the present contents of GW, and is thereby broadcast to the three cortical association areas AC1a to AC3a. The contents of these areas stimulates the association areas AC1b to AC3b to take on patterns of activation corresponding to the expected outcomes of the actions recommended by their motor-cortical counterparts. These

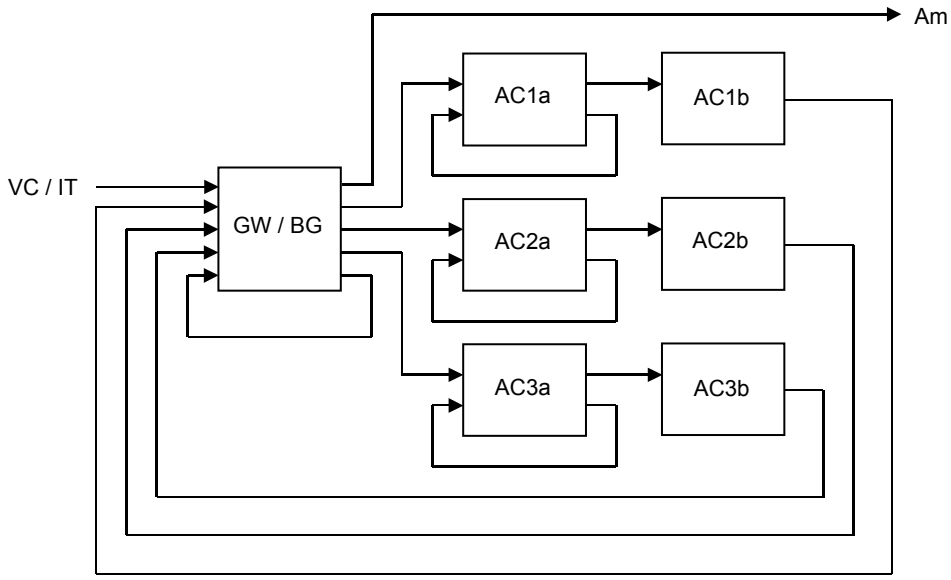

Fig. 9: Circuitry for broadcast and inner rehearsal (GW / BG / AC). GW = global workspace.

patterns are fed back to GW / BG, leading to further associations corresponding to the outcomes of later hypothetical actions. By following chains of associations in this way, the system can explore the potential consequences of its actions prior to their performance, enabling it to anticipate and plan ahead.

But for this capacity to be useful, the system needs to be able to *evaluate* hypothetical futures as it discovers them. So as a result of inner rehearsal, the salience of the currently selected action becomes modulated according to the affective value of the situations to which it might lead (Fig. 7). If the currently selected action potentially leads to a desirable situation, a small population of "reward" neurons becomes active, causing an increase in the salience of that action. This in turn may be sufficient to trigger the release of its veto, bringing about its execution. Conversely, if the currently selected action potentially leads to an undesirable situation, a small population of "punish" neurons becomes active. The resulting decrease in salience of that action may cause a new action to become the most salient. In this case, the transient visual buffer is reloaded, its contents is passed on to GW, and the process of inner rehearsal is restarted, permitting the system to explore a different possible future.

4 Experimental Results

The implemented system currently runs on a 2.5 GHz Pentium 4 machine. Both Webots and NRM are run on the same machine, and the two systems communicate through an internal TCP socket. Under these somewhat unfavourable circumstances, each update cycle for the whole set of neurons takes approximately 750ms. A large proportion of this time is taken up by internal communication and graphics processing.

Fig. 10 illustrates an interesting property of the circuit of Fig. 9. The graph plots the percentage of neurons in the four maps GW and AC1a to AC3a that changed state from one time step to the next during a typical run in which no external sensory input was presented to the robot. (A similar pattern is typically produced soon after the initial presentation of an external stimulus.) The graph shows that the system of inner rehearsal exhibits a procession of stable states punctuated by episodes of instability, a pattern which is reminiscent of the phenomenon of aperiodic alternation between pan-cortical coherent and decoherent EEG activity reported by various authors (Rodriguez, *et al.*, 1999; Freeman & Rogers, 2003).

By computing the mutual information between GW and areas AC1a to AC3a (Fig. 11), it can be shown that the periods of stability depicted in the graph occur when the contents of GW is being successfully broadcast to those three cortical regions. Each peak in Fig. 11 denotes that the corresponding region of association cortex has fallen into a similar attractor to GW. (The absence of simultaneous peaks for all three association areas is due to the differences in their attractor landscapes. When an association area lacks an attractor close to the pattern being broadcast from GW it becomes quiescent.) Conversely, the spikes of instability in Fig. 10 indicate that GW is being nudged out of its previous attractor and is starting to fall into a new one. The new attractor will be the outcome of a competition between AC1b to AC3b. The resulting new contents of GW is then broadcast to AC1a to AC3a, causing new activation patterns to form in AC1b to AC3b, which in turn give rise to a renewed competition for access to GW.

This tendency to chain a series of associations together accounts for the system's ability to look several possible actions into the future. In the way it migrates from one attractor to another, the dynamics of the system can be characterised as *metastable* (Bressler & Kelso, 2001), and its behaviour also bears an interesting resemblance to *chaotic itinerancy* (Tsuda, 2001; Freeman, 2003),

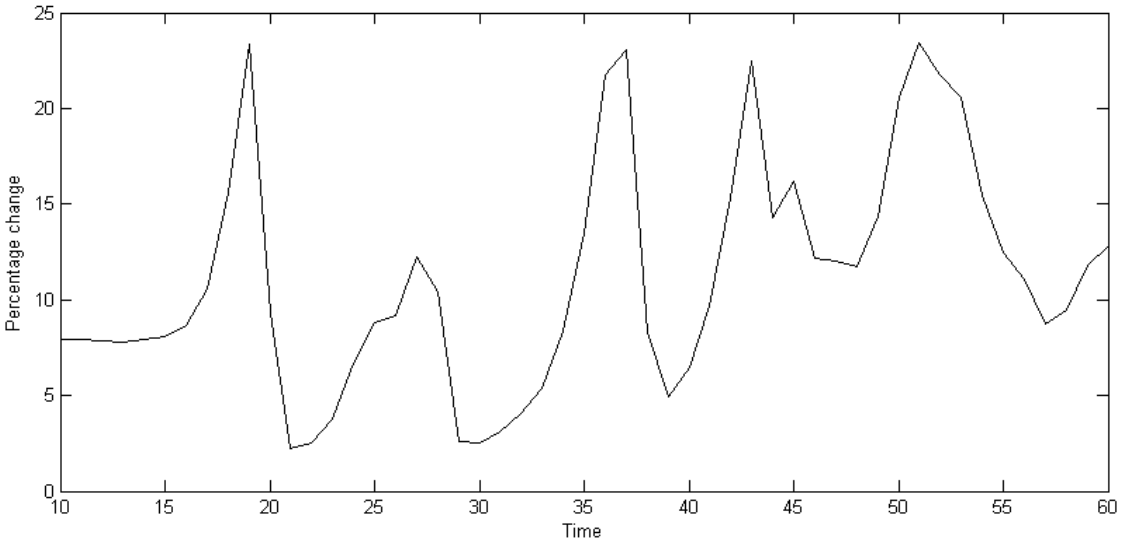

Fig. 10: Cycles of system-wide stability and instability

although in the context of discrete computation the concept is not strictly applicable without some modification.

In these and other respects (such as its use of re-entrant circuitry), the system also echoes Edelman and Tononi's *dynamic core* hypothesis (Edelman & Tononi, 2000; Edelman, 2003; Seth & Baars, 2005), according to which consciousness is realised in the brain by a shifting subset of its neurons (the dynamic core), whose composition is subject to constant and rapid change, but always comprises a set of distributed neural groups whose membership is tightly internally coupled while being functionally separated from the rest of the brain. Similarly, within the present simulation, competing areas of association cortex take turns to dominate the dynamics of the system through access to the global workspace from where their patterns of activation can achieve widespread influence. At any given time, the combination of GW and the areas of association cortex that are migrating to new attractors under its influence (such as AC2a and AC3a at time 20 in Fig. 11) can be likened to the dynamic core in Edelman and Tononi's sense.

Table 1 presents an illustrative sequence of events that occurred in a typical run of the whole system in which this ability to look ahead is put to good use. The episode described starts with the initial presentation of a new stimulus to the robot's camera, and ends with the robot's first action. The time is given in perception-update-action cycles, so the overall time between stimulus and response is around 17 seconds. This suggests that real-time performance would be

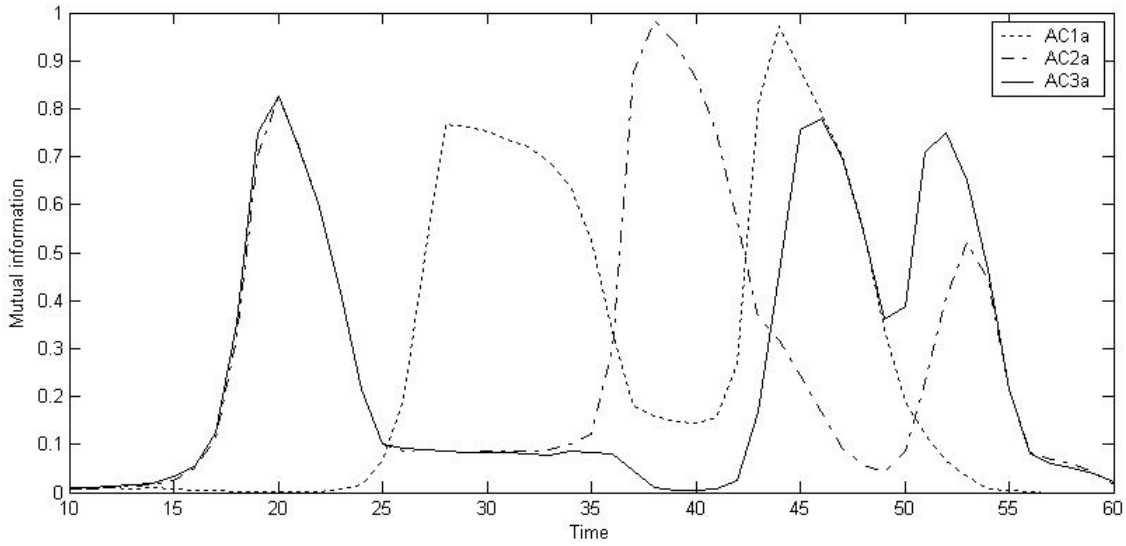

Fig. 11: Mutual information as an index of broadcast

attainable with current technology using a higher-end platform, assuming the Webots simulator is run on a different machine.

For the run presented, the robot's environment contained just three cylinders – one green, one red, and one blue. Area MC1 of the motor-cortical system was trained to recommend "rotate right" (RR) when presented with a green cylinder, while area MC2 was trained to recommend "rotate left" (RL). MC1's recommendation has the higher initial salience, and in a purely reactive system this action would be executed straight away. But thanks to the imposition of a veto, the inner rehearsal system gets a chance to anticipate the outcome of the recommended action. This turns out to be undesirable. So the system considers an alternative action, and this turns out to have a preferable expected outcome so is duly executed.

5 Discussion

Although only a prototype, the implemented system has demonstrated the viability of the proposed architecture. As the episode in Table 1 illustrates, a system conforming to the architecture is capable of generating a cognitively enhanced motor response to an ongoing situation. The design methodology used is, of course, quite different to that currently favoured by researchers in mainstream cognitive robotics (Lespérance, *et al.*, 1994), and is more closely allied to the research programme hinted at by Clark and Grush (1999). In place of viewpoint-free propositional representations, the present system employs viewer-

Table 1: An episode in a typical run

Time	Events
0	Green cylinder comes into view.
4	Green cylinder image in both visual buffers. MC1 recommends RR, MC2 recommends RL. RR has higher salience and is currently selected action. Veto is on.
7	Green cylinder image in GW and broadcast to AC1a to AC3a. AC1b has association with red cylinder, AC2b has association with blue cylinder.
8	Associated red cylinder image now in GW.
11	"Punish" neurons active, salience of RR going down.
13	Salience of RR very low. RL becomes currently selected action.
14	Transient visual buffer reloaded with green cylinder image.
16	Green cylinder image in GW and broadcast to AC1a to AC3a.
20	Associated blue cylinder image now in GW. "Reward" neurons active. Salience of RL going up.
22	Salience of RL very high. Veto released.
23	RL passed on to motor command area. Robot rotates left until blue cylinder in view.

centred analogical representations, and in place of symbolic reasoning it deploys a recurrent cascade of attractor networks. But compared with related products of the classical approach, the current implementation inherits certain several well-known disadvantages.

- While traditional propositional representations possess a compositional structure, and therefore comply with Fodor and Pylyshyn's *systematicity* constraint (Fodor & Pylyshyn, 1988), this is not true of the patterns of neuronal activity in the present system.
- Traditional propositional representations are adept at coping with *incomplete information* using disjunction and existential quantification. The present system can only deal with alternatives by using competitive parallelism and by exploring different threads of possibility at different times.

- Traditional planning systems are typically capable of effecting a *complete search* of the space of possible plans, while the presently implemented system of inner rehearsal ignores large tracts of search space and is only capable of a very crude form of backtracking.

Each of these issues is the subject of ongoing research, and there is a variety of well-known techniques for addressing them. But brain-inspired cognitive architectures are relatively unexplored in artificial intelligence, and much work needs to be done before they can offer a viable alternative to the classical methodology in the domain of cognition.

But in addition to its potential engineering application, the architecture presented here can be construed as a concrete statement of a specific hypothesis about human brain function. In line with the methodological stance outlined in the paper's opening paragraphs, this hypothesis ascribes the capacity for high-level cognition to the interplay of consciousness, emotion, and imagination. Building a computer model and using it to control a robot is one way to give a clear interpretation to these concepts and to make precise their hypothesised role in mediating behaviour.

To conclude, let's consider the extent to which these philosophically difficult concepts of consciousness, emotion, and imagination can legitimately be applied to artefacts that conform to the architectural blueprint of the present paper, such as the implemented robot controller described in the previous section.

Let's begin with the concept of consciousness. The architecture respects all five of the "axioms of consciousness" proposed by Aleksander & Dunmall (2003). However, the present paper draws more heavily on the empirically grounded distinction between conscious and non-conscious information processing hypothesised by global workspace theory (Baars, 1988; 2002). This carries over straightforwardly to the thalamocortical system of Fig. 9. The processing of activation patterns that appear in GW and are subsequently successfully broadcast to cortex can be considered "conscious", while all other information processing that goes on in the system is "non-conscious". In accordance with global workspace theory, information that has been thus processed "consciously" integrates the contributions of many parallel processes, although the parallelism is very small-scale in the implemented robot controller described here.

Similar considerations apply to the concepts of emotion and imagination. The functional role of the affective and inner rehearsal systems in the present

architecture is identical to that proposed for emotion and imagination by many authors for the human case (Damasio, 1995; 2000; Harris, 2000). The argument, in a nutshell, is that "human beings have evolved a planning system in which felt emotion plays a critical role. By imagining what we might do, we can trigger in an anticipatory fashion the emotions that we would feel were we to actually do it" (Harris, 2000, p. 88). In much the same vein, the higher-order loop of Fig. 9 "imagines" what the robot might do, and this triggers an "emotional" response in the affective system of Fig 7.

However, the liberal use of scare quotes in the above paragraphs remains appropriate. There are many possible objections to the literal application of concepts such as consciousness and emotion to a robot such as the one described here. Prominent among these is the sheer poverty of the robot's external environment, the consequent poverty of its control system's internal dynamics, and the limited range of behaviour it can exhibit as a result. But consider a future humanoid robot in an unconstrained natural environment, equipped with a control system conforming to the proposed architecture. Suppose the robot's broadcast / inner rehearsal system comprised not six cortical regions but 100,000. Perhaps it would be harder to rein in the use of these concepts in such a case. But for now this remains pure science fiction.

Acknowledgements

For help, discussion, and inspiration thanks to Igor Aleksander, Bernie Baars, Lola Cañamero, Ron Chrisley, Rodney Cotterill, Yiannis Demiris, Barry Dunmall, Ray Guillery, Gerry Hesslow, Owen Holland, Mercedes Lahnstein, Pete Redgrave, and S.Murray Sherman.

References

Aleksander, I. (1990). Neural Systems Engineering: Towards a Unified Design Discipline? *Computing and Control Engineering Journal 1* (6), 259–265.

Aleksander, I. & Dunmall, B. (2003). Axioms and Tests for the Presence of Minimal Consciousness in Agents. *Journal of Consciousness Studies 10* (4–5), 7–18.

Baars, B.J. (1988). *A Cognitive Theory of Consciousness*. Cambridge University Press.

Baars, B.J. (2002). The Conscious Access Hypothesis: Origins and Recent Evidence. *Trends in Cognitive Science 6* (1), 47–52.

Baars, B.J. & Franklin, S. (2003). How Conscious Experience and Working Memory Interact. *Trends in Cognitive Science 7* (4), 166–172.

Baxter, M.G. & Murray, E.A. The Amygdala and Reward. *Nature Reviews Neuroscience 3*, 563–573.

Bressler, S.L. & Kelso, J.A.S. (2001). Cortical Co-ordination Dynamics and Cognition. *Trends in Cognitive Science 5* (1), 26–36.

Brooks, R.A. (1986). A Robust Layered Control System for a Mobile Robot. *IEEE Journal of Robotics and Automation 2*, 14–23.

Cañamero, L.D. (2003). Designing Emotions for Activity Selection in Autonomous Agents. In R.Trapple, P.Petta & S.Payr (eds.), *Emotions in Humans and Artifacts*, MIT Press, pp. 115–148.

Cardinal, R.N., Parkinson, J.A., Hall, J. & Everitt, B.J. (2002). Emotion and Motivation: The Role of the Amygdala, Ventral Striatum, and Prefrontal Cortex. *Neuroscience and Biobehavioral Reviews 26*, 321–352.

Chrisley, R. (1990). Cognitive Map Construction and Use: A Parallel Distributed Processing Approach. In D.Touretzky, J.Elman, G.Hinton, and T.Sejnowski (eds.), *Connectionist Models: Proceedings of the 1990 Summer School*, Morgan Kaufman, pp. 287-302.

Clark, A. & Grush, R. (1999). Towards a Cognitive Robotics. *Adaptive Behavior 7* (1), 5–16.

Cotterill, R. (1998). *Enchanted Looms: Conscious Networks in Brains and Computers*. Cambridge University Press.

Cotterill, R. (2001). Cooperation of the Basal Ganglia, Cerebellum, Sensory Cerebrum and Hippocampus: Possible Implications for Cognition, Consciousness, Intelligence and Creativity. *Progress in Neurobiology 64*, 1–33.

Damasio, A.R. (1995). *Descartes' Error: Emotion, Reason and the Human Brain*. Picador.

Damasio, A.R. (2000). *The Feeling of What Happens: Body, Emotion and the Making of Consciousness*. Vintage.

Dehaene, S. & Changeaux, J.-P. (2005). Ongoing Spontaneous Activity Controls Access to Consciousness: A Neuronal Model for Inattentional Blindness. *Public Library of Science Biology 3* (5), e141.

Dehaene, S., Sergent, C. & Changeux, J.-P. (2003). A Neuronal Network Model Linking Subjective Reports and Objective Physiological Data During Conscious Perception. *Proceedings of the National Academy of Science 100* (14), 8520–8525.

Demiris, Y. & Hayes, G. (2002). Imitation as a Dual-Route Process Featuring Predictive and Learning Components: a Biologically-Plausible Computational Model. In K.Dautenhahn & C.Nehaniv (eds.), *Imitation in Animals and Artifacts*, MIT Press, pp. 327–361.

Dunmall, B. (2000). *Representing the Sensed World in a Non-Biological Neural System*. Dept. Electrical & Electronic Engineering, Imperial College London.

Edelman, G.M. (2003). Naturlaizing Consciousness: A Theoretical Framework. *Proceedings of the National Academy of Science 100* (9), 5520–5524.

Edelman, G.M. & Gally, J.A. (2001). Degeneracy and Complexity in Biological Systems. *Proceedings of the National Academy of Science 98* (24), 13763–13768.

Edelman, G.M. & Tononi, G. (2000). *A Universe of Consciousness: How Matter Becomes Imagination*. Basic Books.

Fodor, J.A. (2000). *The Mind Doesn't Work That Way*. MIT Press.

Fodor, J.A. & Pylyshyn, Z.W. (1988). Connectionism and Cognitive Architecture: A Critique. *Cognition 28*, 3–71.

Franklin, S. (2003). IDA: A Conscious Artifact? *Journal of Consciousness Studies 10* (4–5), 47–66.

Franklin, S. & Graesser, A. (1999). A Software Agent Model of Consciousness. *Consciousness and Cognition 8*, 285–301.

Freeman, W.J. (2003). Evidence from Human Scalp EEG of Global Chaotic Itinerancy. *Chaos 13* (3), 1–11.

Freeman, W.J. & Rogers, L.J. (2003). A Neurobiological Theory of Meaning in Perception Part V: Multicortical Patterns of Phase Modulation in Gamma EEG. *International Journal of Bifurcation and Chaos 13* (10), 2867–2887.

Fuster, J.M. (1997). *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe*. Lippincott-Raven.

Glasgow, J., Narayanan, N.H. & Chandrasekaran, B. (1995). *Diagrammatic Reasoning: Cognitive and Computational Perspectives*. MIT Press.

Grush, R. (2004). The Emulation Theory of Representation: Motor Control, Imagery, and Perception. *Behavioral and Brain Sciences 27*, 377–396.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D 42*, 335–346.

Harris, P.L. (2000). *The Work of the Imagination*. Blackwell.

Hesslow, G. (2002). Conscious Thought as Simulation of Behaviour and Perception. *Trends in Cognitive Science 6* (6), 242–247.

Hoffmann, H. & Möller, R. (2004). Action Selection and Mental Transformation Based on a Chain of Forward Models. In *Proc. 8th International Conference on the Simulation of Behaviour (SAB 04)*, pp. 213–222.

Holland, O. (2003). Robots with Internal Models. *Journal of Consciousness Studies 10* (4–5), 77–109.

Koch, C. (2004). *The Quest for Consciousness*. Roberts and Company.

Lespérance, Y., Levesque, H.J., Lin, F., Marcu, D., Reiter, R. & Scherl, R.B. (1994). A logical approach to high-level robot programming: A progress report. In B.Kuipers (ed.), *Control of the Physical World by Intelligent Systems: Papers from the 1994 AAAI Fall Symposium*, pp. 79–85.

Lockwood, G.G. & Aleksander, I (2003). Predicting the Behaviour of G-RAM Networks. *Neural Networks 16*, 91–100.

Michel, O. (2004). Webots: Professional Mobile Robot Simulation. *International Journal of Advanced Robotics Systems 1* (1), 39–42.

Mink, J.W. (1996). The Basal Ganglia: Focused Selection and Inhibition of Competing Motor Programs. *Progress in Neurobiology 50*, 381–425.

Nolte, J. (2002). *The Human Brain: An Introduction to its Functional Anatomy*. Mosby.

Picard, R. (1997). *Affective Computing*. MIT Press.

Prescott, T.J., Redgrave, P. & Gurney, K. (1999). Layered Control Architectures in Robots and Vertebrates. *Adapative Behavior 7*, 99–127.

Redgrave, P., Prescott, T.J. & Gurney, K. (1999). The Basal Ganglia: A Vertebrate Solution to the Selection Problem. *Neuroscience 89* (4), 1009–1023.

Rodriguez, E., George, N., Lachaux, J.-P., Martinerie, J., Renault, B. & Varela, F. (1999). Perception's Shadow: Long-Distance Synchronization of Human Brain Activity. *Nature 397*, 430–433.

Seth, A.K. & Baars, B.J. (2005). Neural Darwinism and Consciousness. *Consciousness and Cognition 14*, 140–168.

Shanahan, M.P. (2004). An Attempt to Formalise a Non-Trivial Benchmark Problem in Common Sense Reasoning, *Artificial Intelligence 153*, 141–165.

Shanahan, M.P. (2005). Perception as Abduction: Turning Sensor Data into Meaningful Representation, *Cognitive Science 29*, 109–140.

Shanahan, M.P. & Baars, B. (2005). Applying Global Workspace Theory to the Frame Problem. *Cognition*, in press.

Sherman, S.M. & Guillery, R.W. (2001). *Exploring the Thalamus*. Academic Press.

Sherman, S.M. & Guillery, R.W. (2002). The Role of Thalamus in the Flow of Information to Cortex. *Philosophical Transactions of the Royal Society B 357*, 1695–1708.

Shipp, S. (2003). The Functional Logic of Cortico-pulvinar Connections. *Philosophical Transactions of the Royal Society B 358*, 1605–1624.

Sloman, A. (1971). Interactions Between Philosophy and Artificial Intelligence: The Role of Intuition and Non-Logical Reasoning in Intelligence. *Artificial Intelligence 2*, 209–225.

Sloman, A. (2001). Beyond Shallow Models of Emotion. *Cognitive Processing 2* (1), 177–198.

Stein, L.A. (1995). Imagination and Situated Cognition. In K.M.Ford, C.Glymour & P.J.Hayes (eds.), *Android Epistemology*, MIT Press, pp. 167–182.

Tsuda, I. (2001). Toward an Interpretation of Dynamic Neural Activity in Terms of Chaotic Dynamical Systems. *Behavioral and Brain Sciences 24*, 793–810.

Van Essen, D.C. (2005). Cortico-cortical and Thalamo-cortical Information Flow in the Primate Visual System. *Progress in Brain Research*, in press.

Van Gelder, T. (1997). The Dynamical Hypothesis in Cognitive Science. *Behavioral and Brain Sciences 21* (5), 615–628.

Wolpert, D.M., Doya, K. & Kawato, M. (2003). A Unifying Computational Framework for Motor Control and Social Interaction. *Philosophical Transactions of the Royal Society B 358*, 593–602.

Ziemke, T., Jirenhed, D.-A. & Hesslow, G. (2005). Internal Simulation of Perception: A Minimal Neuro-robotic Model. *Neurocomputing*, in press.