A review of cognitive architectures

ISO project report

Roger Kingdon
MAC 2008-09

Abstract

This paper extends Vernon, Metta and Sandini's review of cognitive architectures by reassessing the designs against five new 'structural' criteria. This approach involves analysing the flowchart representation of each architecture to establish: whether its modules are well-defined; whether its internal connections are suitable; whether there is a double loop for higher-level processing; whether the architecture has been demonstrated in a practical application; and whether its modules map to the main regions of the human brain. By these criteria, the architectures judged to have the greatest promise are ACT-R, Global Workspace, and SOAR. It is acknowledged, however, that none of the reviewed architectures fully address the last criterion in the above list; and this establishes the need for a broader survey, to include more novel but less-established alternative designs.

Contents

1. Introduction

In this paper I review a selection of published cognitive architectures in order to identify the more promising candidates for the simulation of general intelligence in a computer. This review consolidates and extends the recent survey of Vernon, Metta and Sandini (2007) – hereafter abbreviated 'VMS' – by focussing on the *flowcharts* drawn for each of the architectures by their authors. The object here is to complement VMS's *functional* assessment with a second perspective using *structural* criteria. Here, 'structural' is taken to refer to the identity and connectivity of the modules that make up each cognitive design. It follows that a structural assessment of cognitive architectures is best served by directly inspecting the diagrams defining those architectures, that is, their flowcharts (loosely interpreted as diagrams depicting the flow of data and/or control). The utility of this fresh perspective should become increasingly apparent as the review proceeds.

In the remainder of this introductory section I give a brief account of VMS's functional assessment before setting out five new criteria reflecting the structural approach. Section 2 comprises a systematic assessment of the flowcharts of each of the cognitive architectures identified by VMS. A summary of this assessment, and an associated cross-comparison of architectures, appears in Section 3. Finally, Section 4 presents the conclusions of this review.

Architecture assessment criteria

VMS have assessed fourteen cognitive architectures using seven functional criteria, as follows:

Architecture	Paradigm	Embodiment	Perception	Action	Anticipation	Adaptation	Motivation	Autonomy
Soar	C				+	+		
Epic	C		+	+	+			
ACT-R	C		+	+	+	+		
ICARUS	C		+	+	+	+		
ADAPT	C	×	×	×	+	+		
AAR	E	×	×	×			+	×
Global Workspace	E	+	+	+	×		×	×
I-C SDAL	E	+	+	+	+	+	×	×
SASE	E	×	×	×	+	×	×	×
Darwin	E	×	×	+		×	×	×
HUMANOID	H	×	×	×	×	+	+	
Cerebus	H	×	×	×	+	+		
Cog: Theory of Mind	H	×	×	×	+			
Kismet	H	×	×	×			×	

Figure 1.1 Cognitive architectures assessed using seven functional criteria. Copied from Vernon, Metta and Sandini (2007).

In this table each of the architectures is given a rating for 'Embodiment', 'Perception', 'Action', 'Anticipation', 'Adaptation', 'Motivation', and 'Autonomy'. Ratings are: '**x**', indicating that the functional characteristic is 'strongly addressed' in the architecture; '**+**', indicating that it is 'weakly addressed'; and an empty space, indicating that it is not addressed at all. In addition, architectures are categorised as 'cognitivist' ('C'), 'emergent' ('E'), or 'hybrid' ('H'). Each of these 'paradigms' is uniquely defined in terms of five more functional characteristics: 'Computation Operation', 'Representational Framework', 'Semantic Grounding', 'Temporal Constraints', and 'Interagent Epistemology'. The overall impression is of a systematic and quite comprehensive comparative analysis of the cognitive architectures under review. On closer inspection the picture is less straightforward, however. VMS hold back from declaring any particular

architecture to be the outright 'winner' of their review, but, nevertheless, this is what we wish to know. Perhaps the winning architecture is the one that addresses all of VMS's functional criteria to some degree. By this measure, I-C SDAL and SASE are the winners. But I-C SDAL is a rare example of an architecture that has not been implemented in any relevant practical application, whilst the SASE design is cartoon-like in its simplicity (see Figure 2.8). How, then, are we to judge between these architectures? VMS's functional criteria are useful but they are not sufficiently discerning. Instead, I suggest a more direct approach, best summed up by the classic 'engineering' challenge: 'Show me what you're doing; show me how you're doing it.' That is, our review should focus on the operational *system* and its practical *implementation*. In our case, this means looking at the detailed structures of the individual architectures in the context of their proposed or demonstrated applications.

To identify specific 'structural' assessment criteria, a convenient starting-point is the observation that the object of the exercise is to review *cognitive* architectures. That is, it seems reasonable to suppose that the architectures in question relate in some way to the physical structure of the human brain. Taking this psychologist's perspective, Uttal (2001, p.xiii) is particularly scathing:

> The tendency for highly specific microtheories and narrowly constrained findings to proliferate in psychology is as understandable as it is pervasive and bewildering. The subject matter of this science is far more complex and multivariate than perhaps any other. Orderly taxonomies are rare; universal or even broad-ranging theories are nearly non-existent. Indeed, those who aspire to such universality (*e.g*., Newell, 1990; Anderson and Lebiere, 1998)[1] are more likely to offer programming or strategic approaches or descriptive simulations than explanatory 'theories' of psychological processes.

Uttal concludes (p.212) that 'Contemporary efforts to localize psychological functions in specific regions of the brain, particularly when applying the imaging tools, are based on a string of assumptions, many of which are controversial and some of which are demonstrably wrong.' He substantiates this statement by identifying 29 assumptions underlying the 'localization hypothesis' and juxtaposing these with 36 counterassumptions that he associates with his own 'behaviorist' perspective. Central to Uttal's criticism is his argument that whilst the localization hypothesis assumes that 'Brain processes interact in simple ways and can be, to a valid first approximation, described collectively as a linear system by superposition and other arithmetic procedures, as well as by statistical analyses,' in fact this assumption is wrong because 'The brain is a nonlinear system, to which the tools of linear mathematics do not apply.' However, in Kingdon (2009) I set out a novel analysis showing that discrete functional modules interacting with one another in a simple (*i.e*. linear) fashion can generate characteristic nonlinear responses on the collective system scale, just as long as the individual modules are accessed repeatedly in an *iterative* computation. It is instructive to briefly run through this analysis in the current context, as follows.

As part of their analysis, VMS make reference to Maturana and Varela's (1987) ideograms of autopoietic and operationally closed systems. These diagrams are reminiscent of Zilberstein's (2008) 'optimal metareasoning' approach to anytime planning and bounded rationality, as illustrated in Figure 1.2.

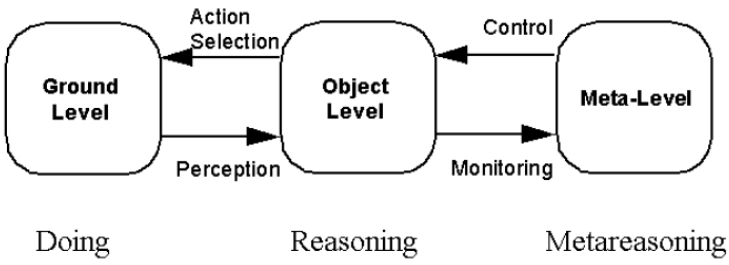

Figure 1.2 The view of metareasoning as monitoring and controlling object-level reasoning. Copied from Zilberstein (2008).

[1] Newell (1990) and Anderson and Lebiere (1998) are reference texts for SOAR and ACT-R respectively.

The significant feature of Zilberstein's design is the 'double loop' created by the addition of a 'metareasoning' module. Figure 1.3 is another version of this system, with precisely the same modules and connectivity, but with a different interpretation given to the agent object-level processor: in this case its role is to compare incoming percepts with internal model predictions generated at the meta-level. In essence this diagram represents primary cognition as a process of *iterative comparison*.

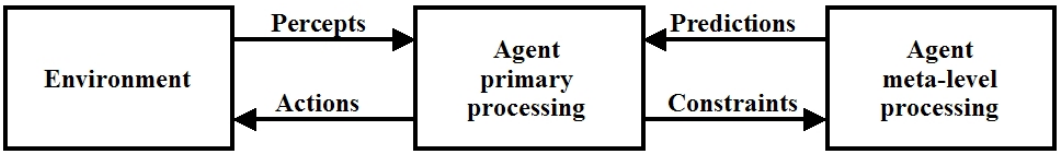

Figure 1.3 Zilberstein's agent architecture reinterpreted as a double processing loop in which internal model constraints are generated through the iterative comparison of incoming percepts with internal model predictions (from the preceding timestep).

From an algorithmic point of view, the purpose of iterative comparison is to obtain a convergence between the agent's predictions and the incoming percepts. Once this has been achieved (for a given threshold) then the agent can have a measure of confidence in its predictions and can act accordingly. But how are we to compare incoming percepts with internal model predictions? An obvious approach is to represent both as regions in 'data space' and to find the extent to which these regions overlap (Figure 1.4).

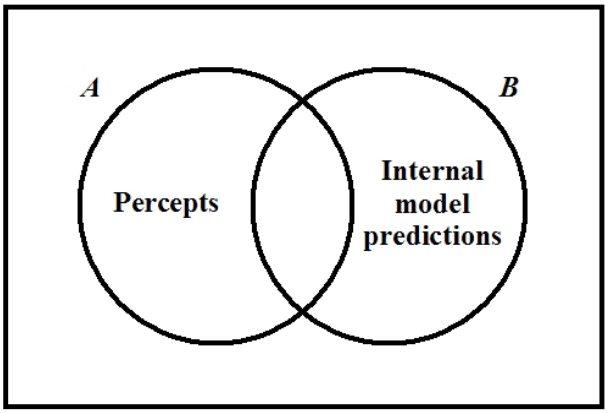

Figure 1.4 The comparison of percepts with internal model predictions, represented as the intersection of two sets A and B in data space.

The Kingdon (2009) analysis now proceeds as follows. First we note that, from its set-theoretic definition, the probability $p(A)$ of arbitrary dataset A is a quantitative measure whose value is proportional to the area occupied by A in data space. Secondly we recall that a standard measure for comparing two datasets is statistical independence, defined as $p(A, B) = p(A)p(B)$ if sets A and B are independent. In this equation $p(A, B)$ is the probability of 'A and B', which is proportional to the area of the intersection of sets A and B. If we define $S(A, B) = p(A, B) - p(A)p(B)$ then we note that, in the general case of statistical *dependence* (that is, if sets A and B are to some extent positively or negatively correlated), $S(A, B)$ will take a non-zero value, and therefore it can be used as a quantitative measure of the 'commonality' of A and B. Now consider the two particular datasets A and B (representing percepts and predictions respectively) in Figure 1.4. As long as the spatial relationship between these datasets does not change 'too much' from one timestep to the next, we can relate their probabilities very simply as $p(A) = \alpha p(B)$ and $p(A, B) = \lambda p(B)$, where α and λ are linear scaling factors. (This amounts to the assumption that linear perturbation theory holds for this application: a common device that is used throughout mathematical physics.) Accordingly we can write $S(A, B) = \lambda p(B)(1 - \alpha p(B)/\lambda)$. Finally we note from Figure 1.3 that on any given timestep the internal model predictions are some function of the internal model constraints, and these constraints are themselves some function of the comparison between incoming percepts and internal model predictions from the preceding timestep: that is, $p(B_{t+1}) = F(A_t, B_t)$ for some composite function F and discrete relative timestep t. If we now choose $F(A, B)$ to be our dataset comparison function $S(A, B)$, and (by way of example) take the simplest case of $\alpha = \lambda$, then we obtain $p(B_{t+1}) = \lambda p(B_t)(1 - p(B_t))$, which is the logistic map in $p(B)$.

Without wanting to labour the point, it is worth noting that the logistic map is a nonlinear iterative equation that has been shown to generate a rich variety of dynamic behaviour, including limit cycles, bifurcations, period doubling, strange attractors, and chaos. Thus this analysis is an effective counter to Uttal's argument that 'The brain is a nonlinear system, to which the tools of linear mathematics do not apply.'

To sum up, this review of cognitive architectures takes a 'structural' approach. This involves inspecting the flowchart representation of each architecture to see whether it 'makes sense': whether its modules are well-defined, whether the module interconnections are suited to the task, and so on. It involves seeing whether there is a double loop that could be used for higher-level processing. It involves asking whether the architecture has been implemented in practice. And, finally, it involves judging the architecture by Uttal's exacting criterion: does it constitute an explanatory theory of psychological processes? Thus we have five new architecture assessment criteria:

1. Are the system modules *well-defined*, that is, have they been assigned distinct processing tasks?
2. Are the connections between the system modules *balanced*, that is, are the connections appropriate, given the specific data and control requirements of the modules that they connect?
3. Does the system have a *double loop* allowing the simulation of higher-level processing?
4. Has the architecture been demonstrated in a relevant practical *application*?
5. Have the system modules been *mapped* to the main regions of the human brain?

These criteria will be revisited in Section 3, following the individual structural assessments of each of the architectures identified by VMS.

2. Review of cognitive architectures

VMS review fourteen cognitive architectures. Using their list notation, these are as follows:

A. SOAR
B. EPIC
C. ACT-R
D. ICARUS
E. ADAPT
F. AAR
G. Global Workspace
H. I-C SDAL
I. SASE
J. Darwin VII
K. Humanoid
L. Cerebus
M. Cog Theory of Mind
N. Kismet

In this section I take another look at these fourteen architectures, but, rather than repeating VMS's *functional* analysis, I choose to focus on the *structural* features of these designs. Accordingly the following review is based on the data and/or control flowchart representations of the architectures, where available.

A. SOAR

The following notes are based on the description of SOAR in Laird, Newell and Rosenbloom (1987), and also on VMS's review.

The SOAR architecture is illustrated in Figure 2.1. The central feature of this design is a double loop structure that implements the alternate execution and update of productions and decisions. Productions are conditional rules (if-then statements), and the set of productions is taken to be a complete representation of the long-term knowledge of the system. These are stored in the 'Production Memory' in the diagram. Decisions are reached through reconciling conflicting statements in 'Working Memory' by means of the 'Decision Procedure' in the lower loop. The operating procedure is as follows:

1. The Working Memory is augmented with declarative statements defining the current state of affairs (for example, the arrangement of objects in Blocks World).
2. The production cycle is invoked whereby the contents of the Working Memory and the Production Memory are compared with one another. Whenever there is a positive match the associated production rule generates a new statement which is added to the Working Memory. This new statement (which can express a preference, a goal or a state of affairs) may in turn generate more positive matches in the next round. This iterative cycle is repeated until there are no more new statements to be added to the Working Memory.
3. Once the production cycle has finished, the decision cycle can begin. The Decision Procedure operates by reconciling the declarative statements that reside in Working Memory as a consequence of previous operations. Reconciliation is effected though the heuristic procedure of 'universal subgoaling' that addresses each 'impasse' between conflicting statements by generating new subgoals, again expressed as statements in Working Memory. These new statements can generate further conflicts which are addressed by repeating the Decision Procedure iteratively (in a similar manner to the production cycle).
4. The successful resolution of an impasse is recorded by the creation of a new rule in the Production Memory by means of the 'Chunking Mechanism'. Essentially this is SOAR's learning mechanism.
5. The system also requires a meta-level 'Working-Memory Manager' that arranges and refines the information in the Working Memory (in particular, deleting statements that have become obsolete). This assists in the automatic switching between the production and decision cycles.

To support their claim that SOAR is an 'Architecture for General Intelligence', Laird, Newell and Rosenbloom identify a broad performance scope ranging from 'typical AI toy tasks' such as Blocks World, through heuristic planning techniques such as iterative deepening, to large-scale expert systems such as NEOMYCIN. By way of example they describe a SOAR implementation of the Eight Puzzle.

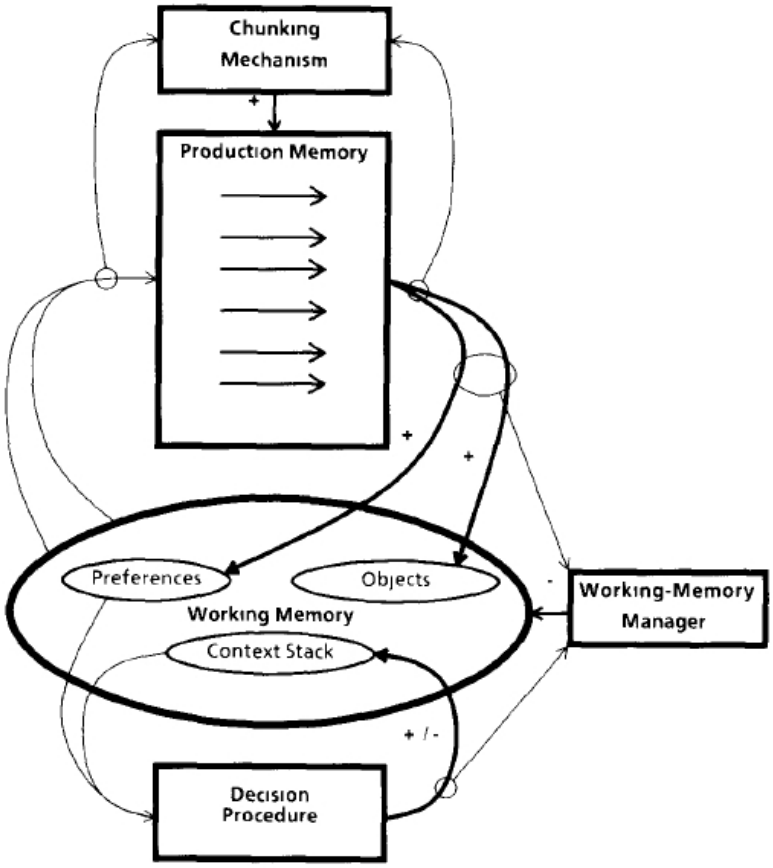

Figure 2.1 The SOAR cognitive architecture. Copied from Laird, Newell and Rosenbloom (1987).

B. EPIC

The following notes are based on the description of EPIC in Kieras and Meyer (1997), and also on VMS's review.

As illustrated in Figure 2.2, EPIC is more a *physical* architecture than a *cognitive* architecture. That is, whilst several input/output (i/o) physical faculties (each with its own processing module) are represented, only two non-i/o cognitive faculties are identified: Memory (divided into 'Production' and 'Long-Term'), and a 'Cognitive Processor' (comprising 'Working Memory' and 'Production Rule Interpreter'). Information flow paths (solid lines) and control connections (dashed lines) interconnect the discrete modules to give a common-sense implementation of a notional anthropomorphic device. The distinction between Working Memory and Production Rule Interpreter in the Cognitive Processor indicates that EPIC is relying on the same declarative approach as SOAR, whilst the absence of information flow paths into Production or Long-Term Memory indicates that EPIC does not have the equivalent of SOAR's learning mechanism.

EPIC provides a high-level framework for the integration of perception/motor modules with a central planning/control unit. As shown in Figure 2.2, the full architecture also includes a simulated 'Task Environment', and this reflects the purpose of the model which is to better understand human-computer interaction (HCI). This involves investing each processor with an operational delay time (estimated from empirical measurements), and ensuring that it can operate independently and in parallel with the other processors and with the Task Environment. Kieras and Meyer describe a number of specific studies undertaken using EPIC, but since these applications focus on HCI rather than cognitive processes (*e.g.* planning or learning), they are not relevant to the current review.

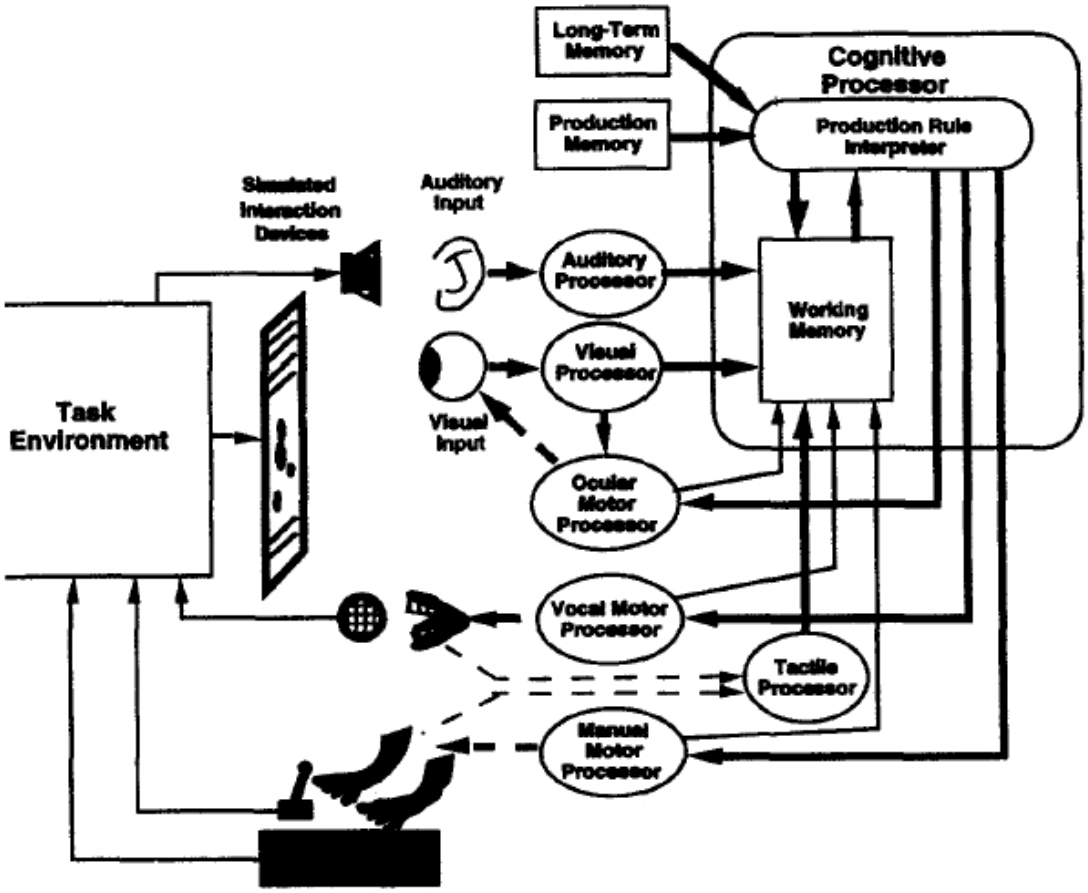

Figure 2.2 The EPIC cognitive architecture. Copied from Kieras and Meyer (1997).

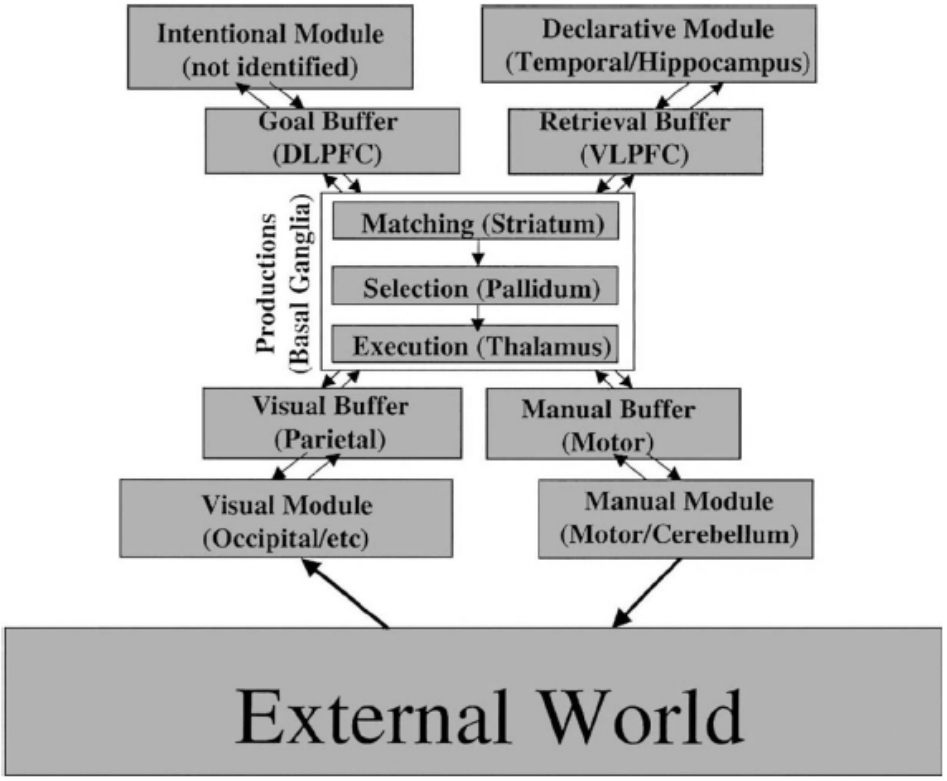

Figure 2.3 The ACT-R cognitive architecture. Copied from Anderson *et al* (2004).

C. ACT-R

The following notes are based on the description of ACT-R in Anderson *et al* (2004), and also on VMS's review.

ACT-R, illustrated in Figure 2.3, has five main modules:

1. The 'Visual Module'. This determines the identity and position of objects in the 'External World'. It corresponds to the occipital lobe at the rear of the cerebrum.
2. The 'Manual Module'. This controls the movements of the hands. Its functionality corresponds to that of the motor cortex at the top of the cerebrum, and the cerebellum tucked beneath the rear of the cerebrum.
3. The 'Declarative Module'. This is the long-term memory store. It corresponds to the temporal lobes at the sides of the cerebrum, and the hippocampus, which is one of the central brain organs enveloped by the cerebrum.
4. The 'Intentional Module'. This keeps track of current goals and intentions. Anderson *et al* have not mapped this module to any specific part of the brain.
5. The 'Central Production System'. This coordinates the activities of the other four modules through information exchange via four shared buffers. These are, respectively, the 'Visual Buffer' (mapped to the parietal lobe at the top rear of the cerebrum), the 'Manual Buffer' (mapped to the motor cortex), the 'Retrieval Buffer' (mapped to the ventrolateral prefrontal cortex, VLPFC), and the 'Goal Buffer' (mapped to the dorsolateral prefrontal cortex, DLPFC). The Central Production System itself is mapped to the striatum, pallidum and thalamus, which reside in the central brain region. Together the striatum and pallidum constitute the basal ganglia at the head of the brain stem.

It is clear from this description that much of the research leading to ACT-R has focussed on analysing the neurological evidence for the existence of the main functional modules and their connections. However, there has also been an effort to relate this model to others having more 'computational' origins. In particular, ACT-R implements the declarative approach and chunking mechanism first seen in SOAR, and it also implements the parallel operation of independent modules as pioneered by EPIC. These features are included by dint of ACT-R's logical and quite general design that strongly resembles a blackboard architecture. This design has the advantage that it can be applied to almost any problem, and the corresponding disadvantage that for any specific problem it is likely to be outperformed by a bespoke heuristic algorithm. Arguably this is just what one should expect from a model which is intended to mimic human cognitive skills, but it does have the undesirable side-effect that Anderson *et al* and associated studies are quite weak when it comes to demonstrating specific applications of ACT-R. Applications do exist, for example, the identification and classification of aircraft tracks on a radar screen, but the feeling remains that the modules in ACT-R are too well-connected for their own good: that is, the model's entirely general connectivity undermines its 'algorithmic muscle'.

D. ICARUS

The following notes are based on the description of ICARUS in Langley (2005), and also on VMS's review.

In their development of SOAR, Laird, Newell and Rosenbloom (1987) set out a number of fundamental modelling principles: the 'physical symbol system hypothesis', the 'goal structure hypothesis', and so on. Curiously, in all these hypotheses there is little or no mention of 'memory', 'decisions', 'objects' or 'preferences', all of which feature strongly in the SOAR architecture (Figure 2.1). It seems that these terms have entered by default, that is, by virtue of the implicit assumption that the modules of a cognitive architecture can be ascribed properties commonly associated with human cognition. One can deal with this assumption either by enforcing it explicitly, as in ACT-R, or by eliminating it altogether, as in ICARUS. Indeed, as VMS explain, ICARUS takes the physical symbol system hypothesis to its logical conclusion by working only with 'symbolic operators' residing in an abstract 'problem space'. Consistent with this approach is the absence of any flowchart representation of ICARUS: clearly, one cannot draw a cognitive architecture if one is calling into question the use of terms such as 'cognitive'; although this logical purity is itself undermined by Langley's repeated reference to ICARUS as a 'cognitive architecture'! In any case, without a flowchart diagram it is not possible to assess ICARUS on the basis of its structural features.

ICARUS has been demonstrated in a number of applications including Blocks World and FreeCell.

E. ADAPT

ADAPT is described by Benjamin, Lyons and Lonsdale (2004) as a 'cognitive architecture specifically designed for robotics'. That is, the focus of ADAPT is on *practical application* rather than the *representation of cognition*. Thus, whilst ADAPT freely avails itself of features already present in SOAR, EPIC and ACT-R, like ICARUS it avoids using the cognitivist terminology. Also like ICARUS, ADAPT is not represented in a flowchart, and this omission places it beyond the scope of this review. ADAPT has been implemented as the control system of a Pioneer P2 robot.

F. AAR

The following notes are based on the description of AAR in Brooks (1986), and also on VMS's review.

The authors of ICARUS and ADAPT are clearly uncomfortable with the 'cognitivist agenda', whereby machine architectures are patterned on human neurological faculties and cognitive processes. Brooks (1986) shares this disquiet, which leads him to reject outright the cognitivist approach in favour of a 'behaviouralist' alternative. Whilst it is difficult to disagree with Christensen and Hooker's (2000a) argument (recounted by VMS) that Brooks' approach is insufficient 'as a principled foundation for a general theory of situated cognition', at least it has the advantage that the resulting design – AAR – is well-furnished with explanatory diagrams (unlike ICARUS and ADAPT). Figure 2.4 is an overview of the AAR design, showing that it comprises a simple autonomous agent having a 'subsumption architecture', that is, multiple levels of control. Brooks explains this architecture by describing levels 0, 1 and 2 of a control system for a mobile robot, see Figure 2.5(a)-(c). Level 0 comprises modules for 'sonar', 'collide', 'feelforce', 'runaway', and 'motor'. These modules and their interconnectivity are unaltered by the addition of level 1 modules for 'wander' and 'avoid'. Likewise, levels 0 and 1 modules and their interconnectivity are unaltered by the addition of level 2 modules for 'grabber', 'pathplan', 'monitor', 'integrate', and 'straighten'. Clearly the operational functionality of the system can be increased further, through the addition of more levels of control. However, for all of the levels the flow of information or control is in the same direction, so that AAR has no equivalent to the SOAR double loop structure. Consequently AAR is fundamentally a reactive system with no capacity to explain self-directed behaviour (as VMS point out).

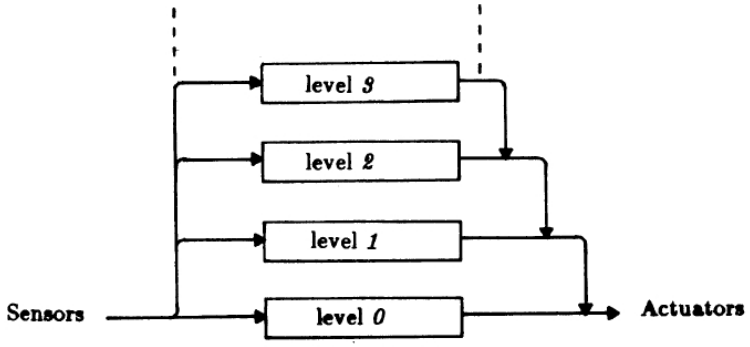

Figure 2.4 Overview of AAR layered control design. Copied from Brooks (1986).

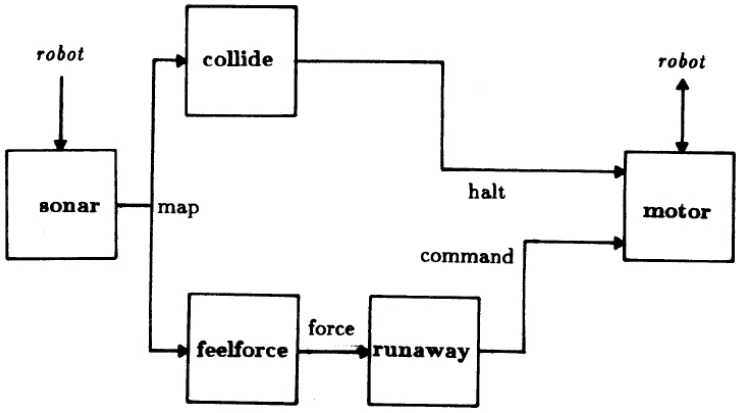

Figure 2.5(a) AAR level 0 for a mobile robot. Copied from Brooks (1986).

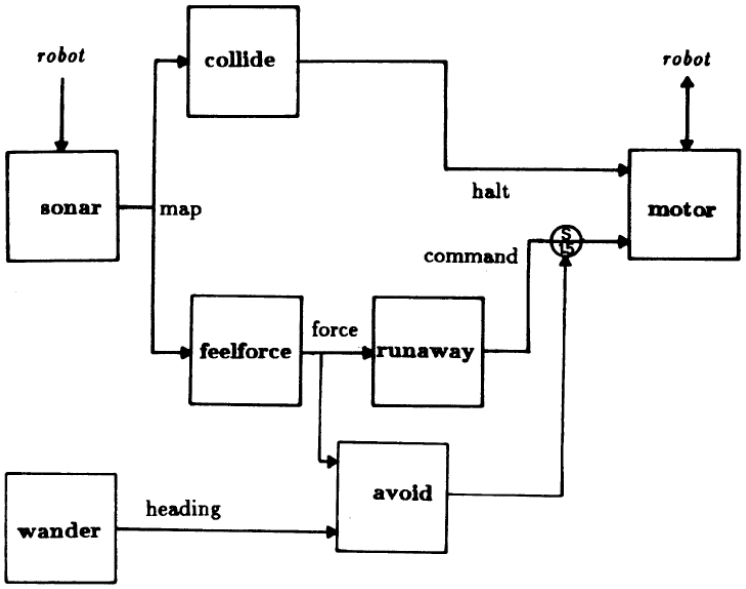

Figure 2.5(b) AAR levels 0 and 1 for a mobile robot. Copied from Brooks (1986).

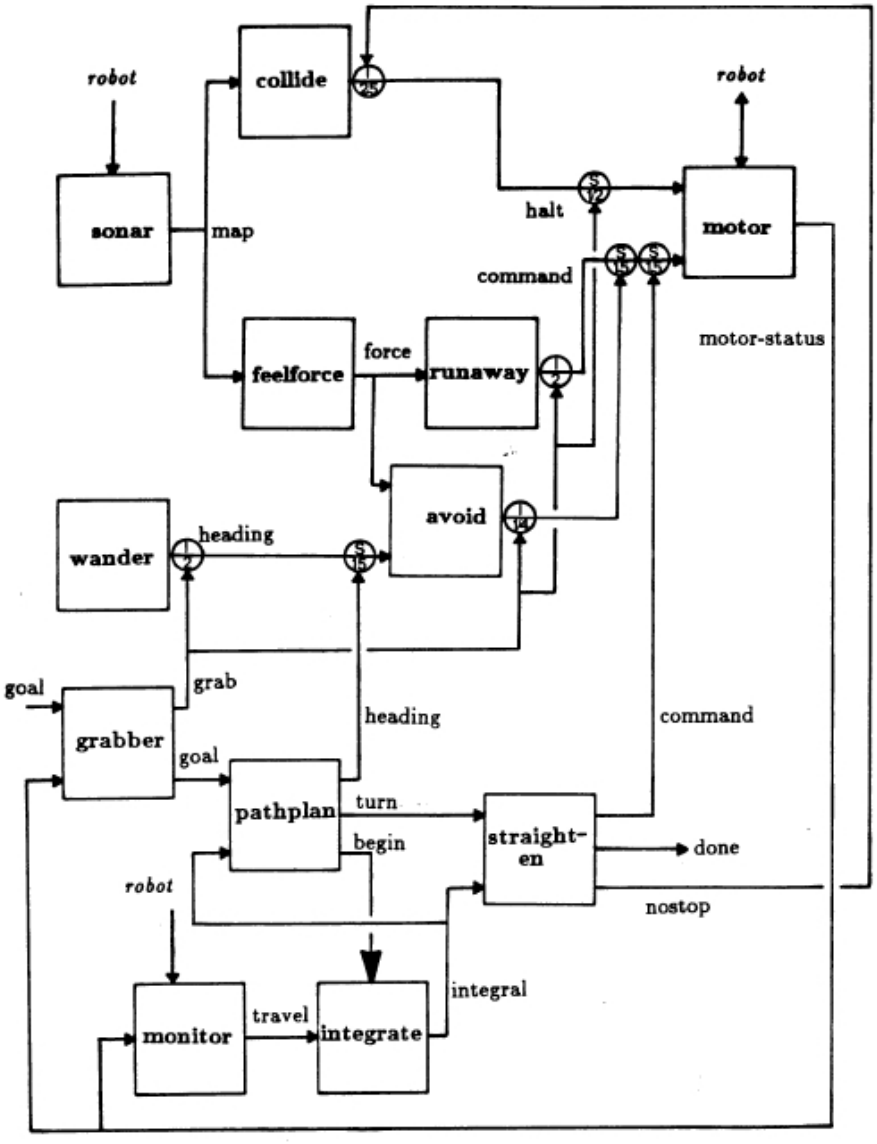

Figure 2.5(c) AAR levels 0, 1 and 2 for a mobile robot. Copied from Brooks (1986).

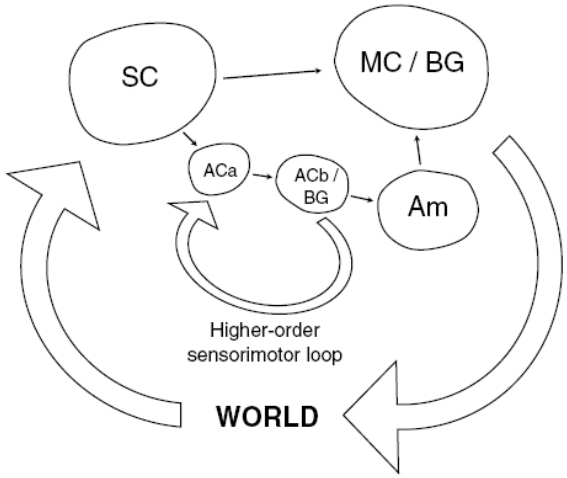

Figure 2.6 A top-level schematic of the Global Workspace cognitive architecture. Copied from Shanahan (2006).

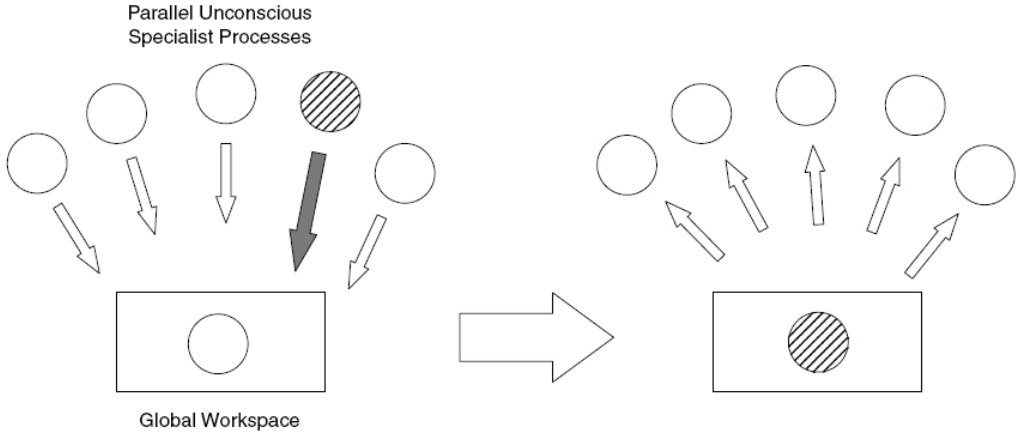

Figure 2.7 Global Workspace parallel processing. Copied from Shanahan (2006).

G. Global Workspace

The following notes are based on the description of the Global Workspace (GW) cognitive architecture in Shanahan (2006), and also on VMS's review.

The GW architecture is predicated on two key observations: (i) The 'simulation hypothesis', that a person's thoughts comprise internal simulations of his interactions with his environment; (ii) The 'global workspace model', whereby the brain's massive parallelism is simulated using a 'global workspace', equivalent to a blackboard architecture. These observations give rise to two complementary mechanisms, illustrated in Figures 2.6 and 2.7 respectively.

Figure 2.6, a top-level schematic of the GW architecture, shows the 'higher-order sensorimotor loop' that implements the simulation hypothesis. Clearly this additional loop is intended to undertake internal processing independent of the immediate promptings of perceptions and planning of actions, so that the resulting architecture is not merely reactive and can – in principle[1] – generate self-directed behaviour. The labels in this diagram identify mappings to parts of the brain: SC, sensory cortex; MC, motor cortex; BG, basal ganglia; AC, association cortex; Am, amygdala. These mappings are not to be taken literally: Shanahan emphasises that they are functional analogues or 'homologies' rather than direct equivalences. Thus they do not alter the main feature of Figure 2.6 which is the additional inner/higher loop that can operate independently of the basic interaction between an agent and its environment.

[1] The reason why it is necessary to include 'in principle' is discussed later, in the review of Cog Theory of Mind.

Figure 2.7 shows the GW mechanism for parallel processing. The circles represent specialist processes that are able to operate independently and in parallel. These processes are able to share information via a global workspace whose role is similar to that of the infrastructure of a communications network. It follows that this workspace is not associated with any anatomically localised region of the brain. Shanahan has implemented this mechanism for the parallel internal simulation associated with the higher-order sensorimotor loop, in application to the control system of a Khepera robot in a simulated Sticks World.

H. I-C SDAL

In their (2000b) Christensen and Hooker outline an interactivist-constructivist model of self-directed anticipative learning (hence I-C SDAL), based on a broad review of cognition theory that takes account of 'neuroethology, psychology, cognitive robotics and philosophy'. However, they do not give a flowchart representation of their model, and neither is it demonstrated in any practical application.

I. SASE

The following notes are based on the description of the SASE cognitive architecture in Weng (2002), and also on VMS's review.

An emergent theme of this review has been the necessity to account for self-directed behaviour by including in the architecture a second independent processing loop (see, in particular, the reviews of AAR and GW). This principle is implemented in its purest form in the SASE architecture, see Figure 2.8. Weng arrives at this architecture by identifying a 'fundamental flaw' in the 'traditional agent' model (Figure 2.9): 'It does not sense its internal 'brain' activities. In other words, its internal decision process is neither a target of its own cognition nor a subject for the agent to explain.' By contrast, the SASE agent 'interacts with not only the external environment but also its own internal (brain) environment: the representation of the brain itself.' Beyond this high-level structure Weng is reluctant to define processing modules and their connections, consistent with his philosophy that a cognitive agent should be capable of learning new tasks *and how to undertake them*. That is, it should be able to develop its own detailed structure according to the task in hand. The SASE architecture has been implemented in the SAIL-2 robotic control algorithm.

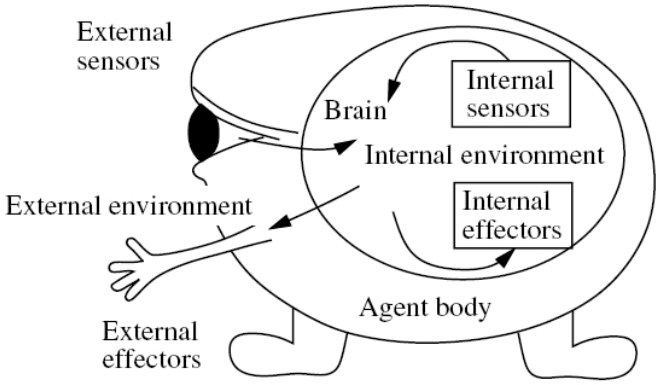

Figure 2.8 The SASE cognitive architecture. Copied from Weng (2002).

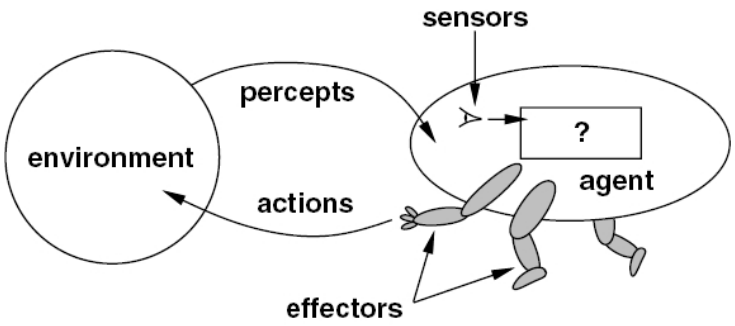

Figure 2.9 The traditional agent architecture. Copied from Weng (2002).

J. Darwin VII

The following notes are based on the description of the Darwin VII cognitive architecture in Krichmar and Edelman (2005), and also on VMS's review.

Darwin VII is one of an evolving series of robot platforms, each with its own bespoke control system. These control systems have been designed following a combined connectionist and structuralist approach, whereby arrays of Hebbian neural units are grouped into a number of different modules, each of which is dedicated to a specific operational task. Figure 2.10 shows the Darwin VII cognitive architecture, which has six main subsystems:

1. Auditory system, comprising modules 'LCoch', 'RCoch' and 'A1'.
2. Visual system, comprising modules 'R', 'VA_PB', 'VA_PH', 'VA_PV' and 'IT'.
3. Taste system, comprising modules 'T_{app}' and 'T_{ave}'.
4. Behaviour (motor response) system, comprising modules 'M_{app}' and 'M_{ave}'.
5. Visual tracking system, comprising module 'C'.
6. Value system, comprising module 'S'.

The numbers in the diagram indicate the quantity of neural units in each module (LCoch, for example, has a 1x64 array of them). System interconnections may be 'excitatory', 'inhibitory', or 'excitatory plastic' for the modification of synaptic strengths. RCoch and LCoch have microphone data inputs; R has a camera data input; T_{app} and T_{ave} have conductivity data inputs; M_{app}, M_{ave} and C have reflex response outputs (to modules R_1, R_2 and R_3 respectively); and S modulates the synaptic strengths of the value-dependent connections (each shown with a black box in the diagram). Modules can operate independently and in parallel, and system learning is achieved through the implementation of synaptic plasticity. In these respects Darwin VII is a clear advance on AAR. On the other hand, for all its sophistication Darwin VII still lacks the double loop present in SOAR, ACT-R, GW and SASE. In operation Darwin VII is capable of exploring a simple environment: avoiding objects, turning towards noise emitters, 'tasting' conducting surfaces, and so on. Later versions have additional functionality, for example, Darwin X (which has 90,000 neural units and 1.4 million synaptic connections) is 'capable of developing spatial and episodic memory', according to VMS. There has been no attempt to map the Darwin modules to specific regions of the brain.

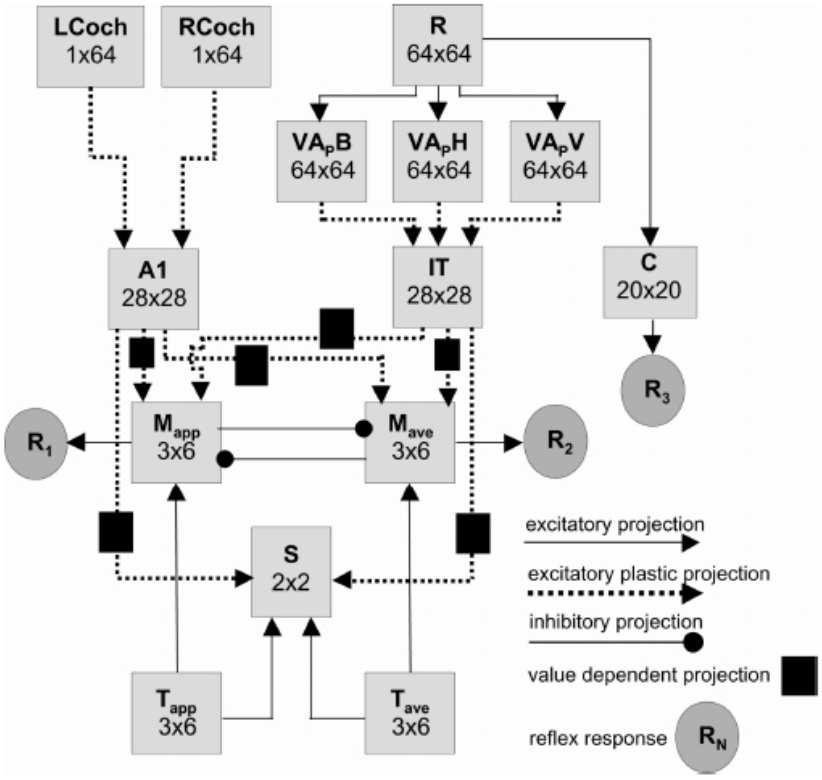

Figure 2.10 The Darwin VII cognitive architecture. Copied from Krichmar and Edelman (2005).

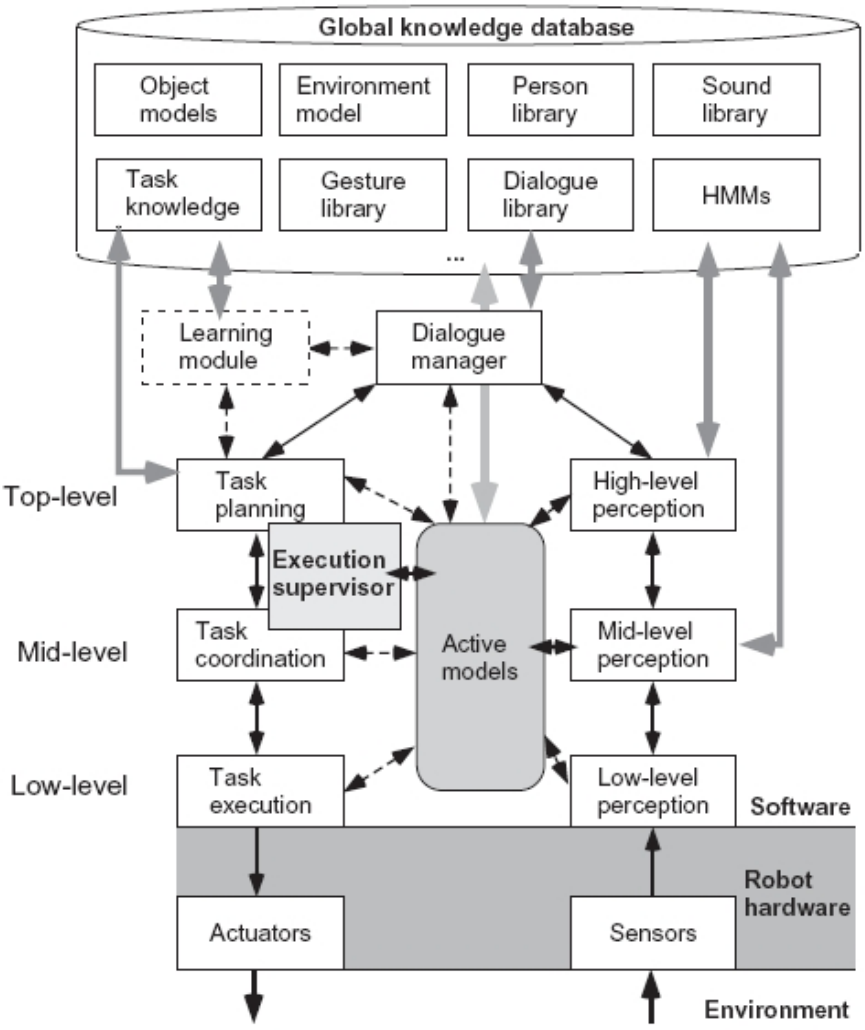

Figure 2.11 The Humanoid cognitive architecture. Copied from Burghart *et al* (2005).

K. Humanoid

The following notes are based on the description of the Humanoid cognitive architecture in Burghart *et al* (2005), and also on VMS's review.

Similar to Darwin, the Humanoid cognitive architecture is a control system for a robot (from which it takes its name). Unlike Darwin, Humanoid does not undertake its processing using neural units, and in this sense Humanoid is closer in character to AAR, albeit with considerably more detail. Indeed, as Figure 2.11 shows, Humanoid has three levels of perception and action, equivalent to AAR's cumulative hierarchy of levels. Humanoid's low, mid, and top levels deal with behavioural sensing/responses, coordination, and planning, respectively. Where appropriate, these modules have access to a 'Global knowledge database' and an 'Active models' unit, comprising long-term memory and working memory, respectively. The Global knowledge database can be updated by a 'Learning module', while the Active models unit is controlled by an 'Execution supervisor'. Perception and action are coordinated at the top level by means of a 'Dialogue manager'. As with AAR and Darwin VII there is no explicit attempt to account for self-directed behaviour by including an additional processing loop, and there is no attempt to map the Humanoid modules to specific regions of the brain.

L. Cerebus

Horswill *et al* (2000) have proposed Cerebus as 'an attempt to extend parallel-reactive architectures to higher-level cognitive tasks'. Like I-C SDAL, Cerebus is a novel departure from established architectures. Also like I-C SDAL, Cerebus is an 'on-going project' that has no flowchart representation and no demonstrated application, which places it outwith this review.

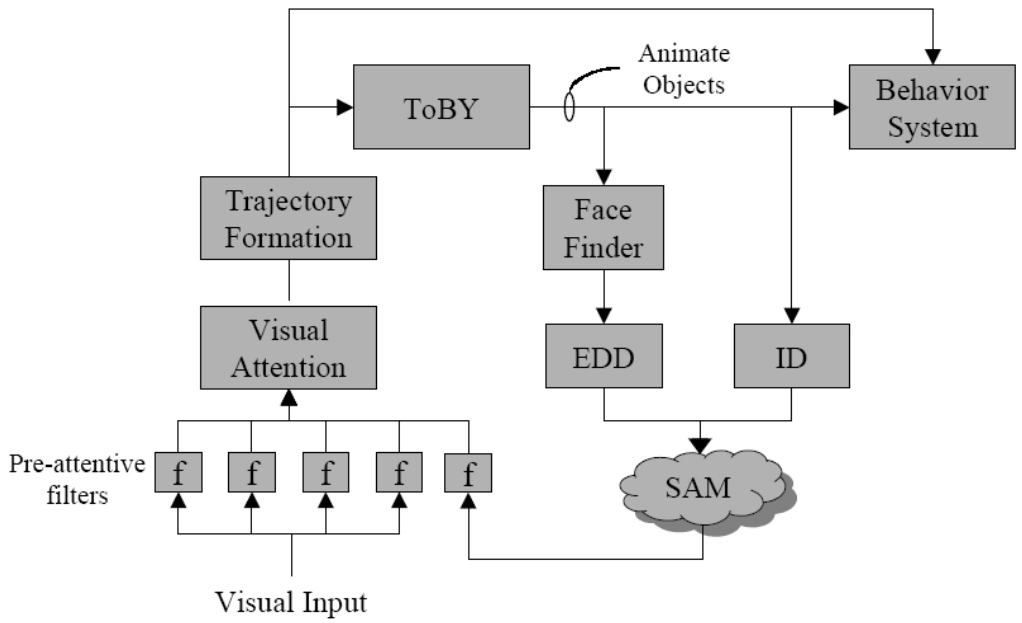

Figure 2.12 An overview of the Cog Theory of Mind cognitive architecture. Copied from Scassellati (2001).

M. Cog Theory of Mind

The following notes are based on the description of the Cog Theory of Mind cognitive architecture in Scassellati (2001), and also on VMS's review.

Like Darwin and Humanoid, Cog is a robot platform intended as a testbed for different cognitive architectures. One of these architectures, the 'Theory of Mind' (ToM) proposed by Scassellati (2001), is included in VMS's review. An overview of this architecture is shown in Figure 2.12. The modules in this diagram are as follows:

1. 'Visual Attention', which handles visual input data via a number of pre-attentive filters 'f'.
2. 'Trajectory Formation', for tracking objects in the visual field.
3. 'Theory of Body' ('ToBY'), for undertaking mechanical actions.
4. 'Behavior System', for the determination of behavioural responses (stance, expression and so on).
5. 'Theory of Mind Mechanism' ('ToMM'), based on Baron-Cohen's (1995) Theory of Mind, comprising 'Face Finder', 'Eye Detection Detector' ('EDD'), 'Intentionality Detector' ('ID') and 'Shared Attention Mechanism' ('SAM').

As shown in the diagram, these modules are connected in a continuous loop that – in principle – corresponds to the additional loop identified in SOAR, ACT-R, GW and SASE, but which is missing from Darwin VII and Humanoid. The 'in principle' caveat is important because it is clear from Scassellati's account that this loop has been included just so that Cog can display a range of 'human-like' behavioural responses, which is a significantly less ambitious purpose than that envisaged by the authors of SASE (in particular). Accordingly it is appropriate to place Cog ToM in the 'middle ground' between those models that include a double loop for the simulation of higher-level cognitive processes (such as SASE) and those models that do not (such as Darwin VII). (It is to be noted that this qualification applies also to the Global Workspace architecture, for a similar reason.) Scassellati does not attempt to map the Cog ToM modules to specific regions of the brain.

N. Kismet

The following notes are based on the description of the Kismet cognitive architecture in Breazeal (2003), and also on VMS's review.

Kismet is a robotic head intended to interact 'face-to-face' with people by perceiving and communicating a wide range of physical and emotional states in real time. Kismet's architecture, illustrated in Figure 2.13,

has five main components:

1. A perceptual system. This comprises several percept handling modules ('Low level Feature Extraction', 'Affective Speech Recognizer', 'Visual Attention', and 'Post-attentive Vision'), and a 'High Level Perceptual System' which stores 'releaser processes' (roughly equivalent to SOAR's productions) expressing Kismet's current set of beliefs about itself and its environment.
2. The 'Emotion System'. Here, each state-description generated by the perceptual system is tagged with arousal, valance and stance ('A, V, S') affects. These affects then elicit different emotions whose strengths are evaluated over all active states. Finally in an 'Emotion Arbitration' phase a single emotion is selected as representative. This dominant emotion is then communicated to the behaviour and motor systems, and the associated affects are communicated to the perceptual and motor systems.
3. The 'Behavior System'. This determines Kismet's behavioural response to its current state, modulated by its dominant emotion and its drives.
4. A set of 'Drives'. These express Kismet's top-level goals of social interaction, play, and rest.
5. The 'Motor System'. This controls Kismet's actions.

At the heart of this design are the releasers, which (as VMS point out) are individually 'handcrafted' by the system designer. The similarity of releasers and productions has been noted already. It is sobering to reflect that, whereas it was always hoped that cognitive systems such as SOAR would be able to write their own productions, the experience with Kismet has shown that practical implementation involves very little in the way of machine self-learning. That is, whilst Kismet has an additional closed loop (passing the affective state back to the High Level Perceptual System) that 'in principle' delivers the internal autonomy required for self-direction or self-learning, in practice this mechanism has not yet bypassed the human-in-the-loop.

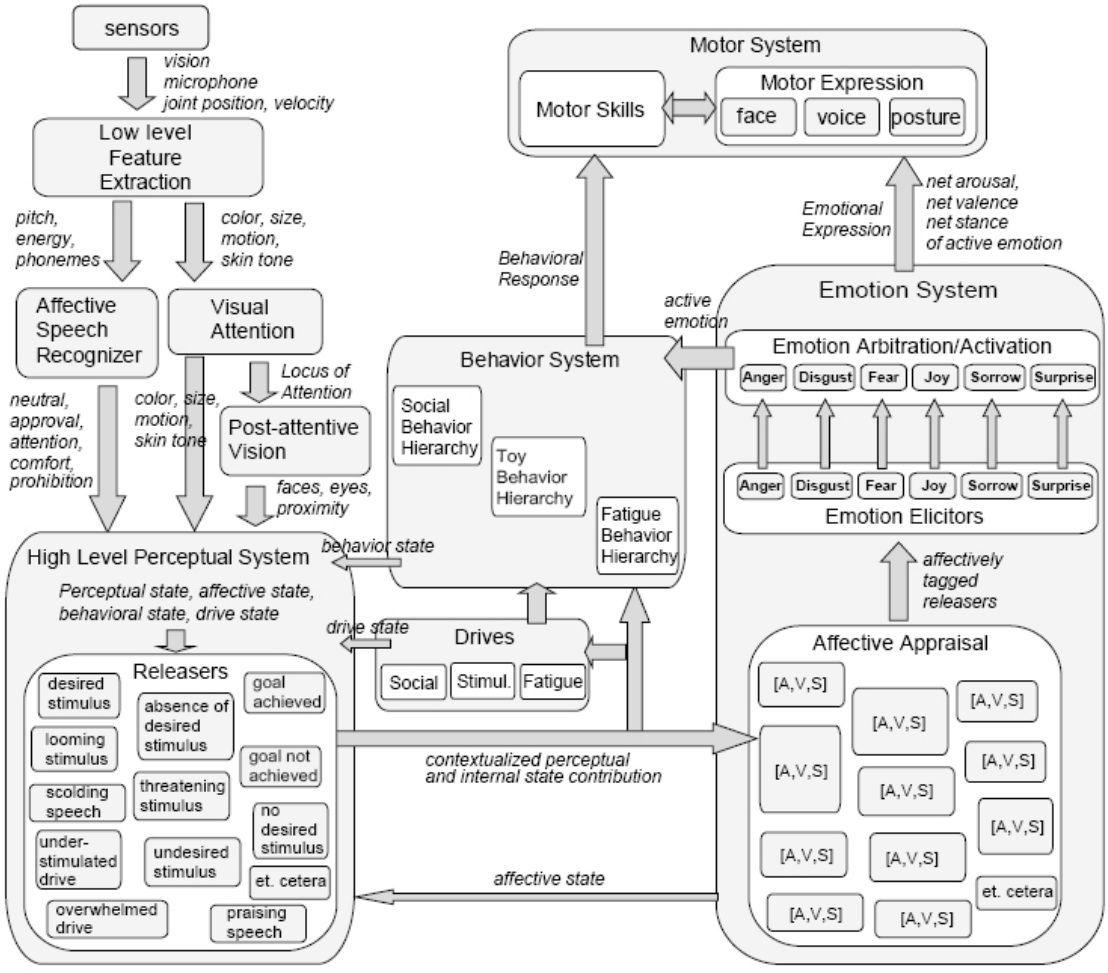

Figure 2.13 The Kismet cognitive architecture. Copied from Breazeal (2003).

3. Discussion

Summary and cross-comparison

As VMS have shown, it is possible to obtain a succinct summary and cross-comparison of the reviewed cognitive architectures by constructing a 'balanced scorecard', that is, by tabulating the performance of each architecture against each of the selected assessment criteria. Using the five structural assessment criteria identified in Section 1, we obtain the following:

Model	Well-defined modules	Balanced connectivity	Double loop	Relevant application	Mapped to brain	Total ✓
SOAR	✓	✓	✓	✓	×	4
EPIC	×	✓	×	×	×	1
ACT-R	✓	×	✓	✓	✓	4
ICARUS	?	?	?	✓	×	1
ADAPT	?	?	?	✓	×	1
AAR	✓	✓	×	✓	×	3
GW	✓	✓	✓/×	✓	✓/×	4
I-C SDAL	?	?	?	×	✓/×	½
SASE	×	✓	✓	✓	×	3
Darwin VII	✓	✓	×	✓	×	3
Humanoid	✓	✓	×	✓	×	3
Cerebus	?	?	?	×	×	0
Cog ToM	✓	✓	✓/×	✓	×	3½
Kismet	✓	✓	✓/×	✓	×	3½

Figure 3.1 Summary assessment of the fourteen cognitive architectures.

In this table the ratings are: '✓', indicating that the criterion is addressed in the architecture (roughly equivalent to VMS's 'x'); '✓/×', indicating ambivalence (roughly equivalent to VMS's '+'); '×', indicating that the criterion is not addressed (equivalent to VMS's empty space); and '?', indicating that no information is available. The final column gives an indicative total for each architecture, with ✓ scoring 1 and ✓/× scoring ½.

Focussing mainly on the 'double loop' column, it is possible to identify four different design categories:

1. Architectures that have a double loop for the simulation of higher-level cognitive processes. These are SOAR, ACT-R, and SASE. Mean score 3⅔.
2. Architectures that have a double loop for the simulation of behavioural processes. These are GW, Cog ToM, and Kismet. Mean score 3⅔.
3. Architectures that do not have a double loop. These 'reactive' systems are AAR, Darwin VII, and Humanoid. Mean score 3.0.
4. Designs that are not *demonstrated cognitive architectures*. These are EPIC, ICARUS, ADAPT, I-C SDAL, and Cerebus. Mean score 0.7.

This categorisation is insightful, up to a point: whilst mean score clearly separates the last category from the rest, it draws no distinction between the first two categories. Another approach is to rank architectures just by their individual totals:

1. Total score 4: SOAR, ACT-R, GW.
2. Total score 3½: Cog ToM, Kismet.
3. Total score 3: AAR, SASE, Darwin VII, Humanoid.
4. Total score 1 or less: EPIC, ICARUS, ADAPT, I-C SDAL, Cerebus.

This ranking confirms the joint-pole position of SOAR and ACT-R, and promotes GW at the expense of SASE, which drops out of the running (to borrow a couple of well-known sporting phrases). It seems, however, that there is no clear 'winner', unless we take a more radical approach and refer exclusively to the 'mapped to brain' criterion, where only ACT-R stands out. Or maybe GW is the winning architecture, because it alone addresses all of the criteria to some extent. Or maybe we should construct a weighted combination of these structural criteria together with VMS's functional criteria, but this greatly detracts from the clarity and simplicity of the summary tables. In any case it is clear that, whilst both Figures 1.1 and 3.1 offer useful general insights, it is also necessary to identify particular features of architectures that can make-or-break their candidacy. To which end, the following specific comments[1] are of relevance:

- There does not appear to be any automatic mechanism for the refinement of SOAR's Chunking Mechanism, Working-Memory Manager, or Decision Procedure. Furthermore, it seems that SOAR's impasse-resolution mechanism is nothing more than a reformulation of Hegelian dialectic (where a thesis contradicted by an antithesis may be resolved to give a meaningful synthesis), which has been shown by Popper (1957) to be fatally flawed.
- From a computational point of view, ACT-R's main problem is its excessive connectivity. From a neuroscientific point of view, ACT-R's main problem is that its 'blackboard architecture' has no equivalent in the human brain. This is the underlying reason why Anderson *et al* have been unable to map the Intentional Module to any specific region of the brain.
- GW's two mechanisms (Figures 2.6 and 2.7) are 'complementary' in the sense that they do not invalidate one another. On the other hand, these mechanisms do not necessitate one another: indeed, they appear to be entirely independent. The GW parallel processing mechanism (Figure 2.7) is a blackboard architecture that has no equivalent in the human brain, which is why Shanahan carefully describes his references to brain regions as 'analogues'.

Given these points, if I had to come off the fence and declare a 'winner', I would choose ACT-R, with GW a close second. But in truth the real winner is William Uttal, whose damning assessment of the state-of-the-art is borne out by the proliferation of '×'s in the 'mapped to brain' column of Figure 3.1.

Dénouement

I would not have been so interested in reviewing this leading selection of published cognitive architectures if I did not have my own ideas on the subject. My design, called 'IDEAL', is shown in Figure 3.2.

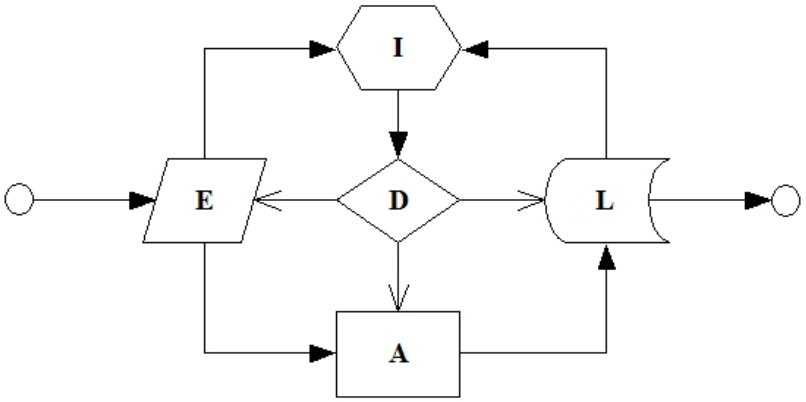

Figure 3.2 The IDEAL cognitive architecture. Copied from Kingdon (2009).

The modules in this diagram are as follows: 'I' handles iteration and data comparison; 'D' handles decision-making; 'E' handles the input of empirical data (percepts); 'A' handles algorithmic processing; and 'L' is a long-term memory log (including the storage of data for fine motor control). The circles on the left and right are connection nodes for input from/output to the environment, respectively. Filled arrow connections pass data, and barbed arrow connections pass control commands.

[1] These comments have their origins in 'marginalia': observations that were made during the review but which were originally deemed irrelevant, in the sense that they did not appear to relate to the defined assessment criteria.

One way to understand IDEAL is to think of it as the result of augmenting a GW-style 'simulation hypothesis' with a novel 'comparison hypothesis'. We recall that the simulation hypothesis asserts that a person's thoughts comprise internal simulations of his interactions with his environment. To this, the proposed comparison hypothesis adds the qualifier that a person's thoughts *also* comprise the identification and recognition of percepts, and this is accomplished by means of data comparison. Simulation and data comparison are undertaken by the A and I modules respectively.

Details of IDEAL's various modes of operation are given in Kingdon (2009). For present purposes it is sufficient to record how IDEAL addresses the current objections to SOAR, ACT-R, and GW. First, the absence of any module having two control inputs means that IDEAL avoids SOAR's pitfall of being compared with Hegelian dialectic. Secondly, in Kingdon (2009) I argue that, whilst it is feasible to construct a blackboard architecture having the same basic functionality as IDEAL, such a system would never be seen in nature because it is not energy-efficient. Figure 3.3 shows this alternative design, which is remarkably similar to the ACT-R cognitive architecture (Figure 2.3). Finally, Figure 3.4 shows the IDEAL modules mapped to the main regions of the human brain: I to the parietal lobe, D to the frontal lobe, E to the occipital lobe, A to the temporal lobes, and L to the cerebellum. This mapping has been obtained by working-up through the brain structures of fish, reptiles, birds and mammals: an approach which even William Uttal might respect as a sound basis for an explanatory theory. Combined, these features give IDEAL a total score of 4 against the structural assessment criteria. The reason why IDEAL does not score a perfect 5 is that it has yet to be implemented in a relevant application; but that may change.

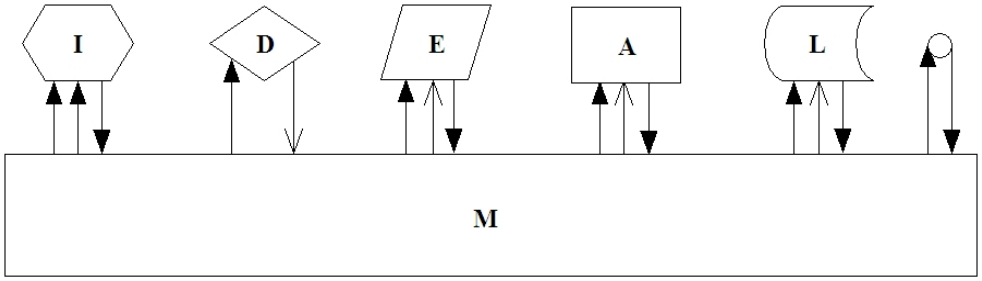

Figure 3.3 Blackboard architecture equivalent to IDEAL. 'M' is a massive internal memory. Copied from Kingdon (2009).

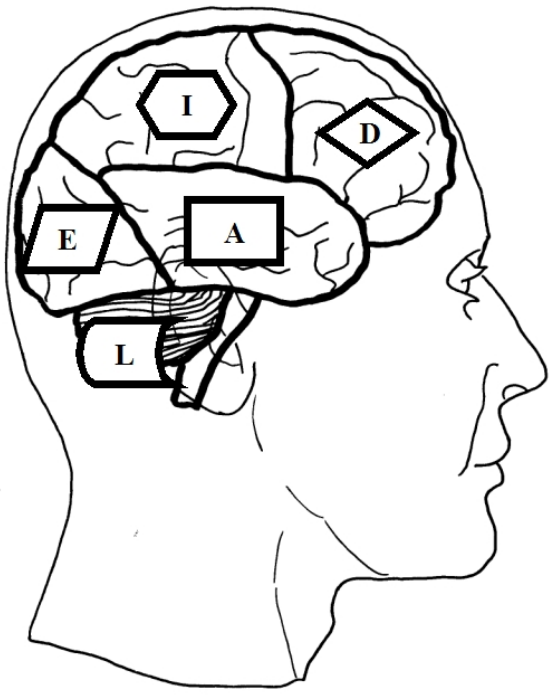

Figure 3.4 IDEAL modules mapped to localised regions of the brain. Copied from Kingdon (2009).

4. Conclusions

By analysing their flowchart designs, it is possible to reassess cognitive architectures against structural criteria. This new approach complements the valuable benchmark review of Vernon, Metta and Sandini, and offers additional and deeper insight. Specifically, the architectures judged by these new criteria to have the most promise are ACT-R, Global Workspace, and SOAR (in that order). However, this review also concedes Uttal's charge that 'universal or even broad-ranging theories are nearly non-existent'. Accordingly it is recommended that researchers look beyond the established state-of-the-art and give consideration to more novel architectures, in particular, my own IDEAL design.

Acknowledgements

This research was conducted under the supervision of Professor Murray Shanahan and it has benefited considerably from his ideas and guidance.

References

Anderson, J. R., and Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., and Qin, Y. (2004). An integrated theory of the mind. *Psy. Rev.*, vol. 111, no. 4, pp. 1036-1060.

Baron-Cohen, S. (1995). *Mindblindness*. Cambridge, MA: MIT Press.

Benjamin, D., Lyons, D. and Lonsdale, D. (2004). ADAPT: A cognitive architecture for robotics. In A. R. Hanson and E. M. Riseman (Eds), *Proc. Int. Conf. Cognitive Modeling*., Pittsburgh, PA.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *Int. J. Human-Computer Studies*., vol. 59, pp. 119-155.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE J. Robot. Autom*., vol. RA-2, no. 1, pp. 14-23.

Burghart, C., Mikut, R., Stiefelhagen, R., Asfour, T., Holzapfel, H., Steinhaus, P., and Dillman, R. (2005). A cognitive architecture for a humanoid robot: A first approach. In *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, pp. 357-362.

Christensen, W. D., and Hooker, C. A. (2000a). Representation and the meaning of life. In *Representation in Mind: New Approaches to Mental Representation*. Sydney, Australia: Univ. Sydney.

Christensen, W. D., and Hooker, C. A. (2000b). An interactivist-constructivist approach to intelligence: Self-directed anticipative learning. *Philosoph. Psy*., vol. 13, no. 1, pp. 5-45.

Horswill, I., Zubek, R., Khoo, A., Le, C., and Nicholson, S. (2000). The Cerebus project. In *AAAI Fall Symposium on Parallel Cognition and Embodied Agents*.

Kieras, D., and Meyer, D. (1997). An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*, vol. 12, no. 4, pp. 391-438.

Kingdon, R. D. (2009). *Principia Intellegentia: The principles governing human and machine intelligence*. New Delhi, India: Allied Publishers.

Krichmar, J. L., and Edelman, G. M. (2005). Brain-based devices for the study of nervous systems and the development of intelligent machines. *Artif. Life*, vol. 11, pp. 63-77.

Laird, J. E., Newell, A., and Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. In S. Sternberg and D. Scarborough (Eds), *Invitation to Cognitive Science vol. 4, Methods, Models, and Conceptual Issues*. Cambridge, MA: MIT Press.

Langley, P. (2005). An adaptive architecture for physical agents. In *Proc. IEEE/WIC/ACM Int. Conf. Intell. Agent Technol.*, Compiegne, France, pp. 18-25.

Maturana, H., and Varela, F. (1987) *The tree of knowledge: The biological roots of human understanding*. London, UK: New Science Library.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Popper, K. R. (1957). *The Poverty of Historicism*. London, UK: Routledge.

Scassellati, B. M. (2001). *Foundations for a theory of mind for a humanoid robot*. PhD thesis, MIT, MA.

Shanahan, M. P. (2006). A cognitive architecture that combines internal simulation with a global workspace. *Consciousness and Cognition*, vol. 15, pp. 433-449.

Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. Cambridge, MA: MIT Press.

Vernon, D., Metta, G., and Sandini, G. (2007). A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE Transactions on Evolutionary Computation*, vol. 11, no. 2, pp. 151-180.

Weng, J. (2002). A theory for mentally developing robots. In *Proc. 2nd Int. Conf. Development and Learning*, pp. 131-140.

Zilberstein, S. (2008). Metareasoning and bounded rationality. *AAAI Workshop on Metareasoning: Thinking about Thinking*, Chicago, IL.