In: *Philosophical Transactions of the Royal Society B*, 362 (1603), pp. 2705–2715.

The Brain's Connective Core and its Role in Animal Cognition

Murray Shanahan
Department of Computing,
Imperial College London,
180 Queen's Gate,
London SW7 2AZ, UK.

Abstract

This paper addresses the question of how the brain of an animal achieves cognitive integration, that is to say how it manages to bring its fullest resources to bear on an ongoing situation. To fully exploit its cognitive resources, whether inherited or acquired through experience, it must be possible for unanticipated coalitions of brain processes to form. This facilitates the novel recombination of the elements of an existing behavioural repertoire, and thereby enables innovation. But in a system comprising massively many anatomically distributed assemblies of neurons, it is far from clear how such open-ended coalition formation is possible. The present paper draws on contemporary findings in brain connectivity and neurodynamics, as well as the literature of artificial intelligence, to outline a possible answer in terms of the brain's most richly connected and topologically central structures, its so-called connective core.

1. Introduction

The brain of an animal is a massively parallel, distributed control system that attempts to maintain certain metabolic variables within acceptable bounds while fulfilling a procreative mission. Evolution has solved a very difficult control problem, namely how to effectively exploit all those parallel, distributed resources when responding to a large variety of unpredictable circumstances. Moreover, since the brain of a bee contains fewer than one million neurons while that of a human contains around 100 billion, evolution has found a solution (or solutions) to this problem on a wide range of scales (Chittka & Niven, 2009). What are the underlying principles of the solution or solutions that evolution has found?

The same question can be recast in more cognitive terms. Let us say that perfect *cognitive integration* is achieved when the animal brings the totality of what it knows to bear on the ongoing situation — its grasp of the sensorimotor contingencies of multiple domains and its understanding of the associated affordances, plus the full contents of both its long-term (episodic-like) memory and its short-term (working) memory. Humans excel at cognitive integration. Yet failures of cognitive integration are commonplace even in humans, such as when I remove the U-bend of my sink then pour its dirty contents back into the plug-hole, causing an unwelcome deluge. On the other hand, apparent triumphs of cognitive integration have been demonstrated in non-human animals, such as innovative tool manufacture (Weir, *et al*., 2002; Bird & Emery, 2009) [1] and sequential tool use (Bird & Emery, 2009; Wimpenny, *et al*., 2009; Taylor, *et al*., 2010). [2] How is cognitive integration achieved in the biological brain?

Drawing on the literature of computer science and artificial intelligence as well as that of neuroscience, this paper surveys a range of mechanisms that might underlie some of the most impressive feats of cognitive integration performed by non-human animals. The focus is on achievements, such as those just cited, that appear to require combinatorial search of a space of possible sequences of action, or (to put it another way) that show signs of "insight" in the sense of Epstein (1985, p.627), who defines insightful behaviour as "The spontaneous interconnection of two repertoires of behavior which [have] been established separately". (For a contemporary discussion of the concept see Shettleworth (2010).) The paper advocates a mix of serial and parallel processing to explain these accomplishments. However, in the context of a massively parallel neural substrate, the serial processing in question bears little resemblance to that of a conventional computer. Rather, each serial step results from the competitive formation of a (possibly novel) coalition of brain processes drawn from a large repertoire of potential combinations. To facilitate this blend of serial and parallel, it is hypothesised that a certain pattern of connectivity is required, wherein

[1] Four-year old children consistently fail at the wire-bending task while eight-year old children typically succeed (Beck, *et al*., 2011), a result which lends credibility to this paradigm as a benchmark for a distinctive capacity for innovation.

[2] These experiments are used here as motivating examples. The aim is not to explain their results *per se*. Attempting to do this without further development of the theory would most likely result in just-so stories. The paper uses examples from avian cognition. For a wider overview of animal tool use, see Seed & Byrne (2011).

information and influence funnels in to and fans out from a connective core of cognitively important brain regions, resulting in an integrated response to the animal's ongoing situation.

2. Explaining Cognitive Prowess in Terms of Computation

We will begin with explanations that inherit their terms from classical artificial intelligence and cognitive science, then move to explanations in the language of neurodynamics. The cognitive feats of interest all feature an element of *planning*, that is to say, the discovery of a novel sequence of actions to achieve a goal. [3] In the sequential tool use paradigm of Bird & Emery (2009), a stone has to be dropped into a narrow tube to release a trap-door. But no suitably-sized stone is available. Instead, the bird has to first drop a large stone into a wide tube to release a small stone, which can then be used in the narrow tube. Similarly, in the wire-bending paradigm of Weir, *et al*. (2002), the bird has to retrieve a bucket from a tube. But the bucket is out of reach, and a tool is required to retrieve it. Moreover, no suitable ready-made tool is available, and success at the task requires a straight piece of wire first to be bent into a hook.

Success in such tasks only counts as planning if it is achieved on first trial, when the subject is unfamiliar with the problem. As soon as the subject has performed the task successfully once, planning is no longer strictly necessary. Moreover, even if the task is performed successfully on first trial, this cannot be counted as planning if it is the result of extensive trial-and-error and / or luck. These explanations for success have to be ruled out. Of course, an innate behaviour involving sequences of actions, such as nest-building (Hansell, 2007), doesn't necessarily require planning either, however complex it might be. The hallmark of a genuine planning task is the need to search a tree of alternative sequences of actions.

The topic of planning has received extensive treatment in the field of artificial intelligence (AI) (Russell & Norvig, 2010, Ch. 10), so it is instructive to consider the range of methods that can be deployed for solving planning problems on a computer while asking whether the brains of clever animals might operate in a similar way. Planning algorithms that are predominantly serial rather than parallel can be classified

[3] The term "planning" is sometimes used in the animal cognition literature to mean planning for the future needs (Raby, *et al*., 2007). Here we are using the term in the sense prevalent in artificial intelligence and robotics.

according to the order in which they explore a state space. A *backward chaining* technique is one that works backwards from a desired goal state, via an action that would achieve that goal state, to a subsidiary goal state. This process is then repeated until a state is reached that matches the current state of the world. The sequence of actions that is built up along the way is a plan. Conversely, a *forward chaining* technique begins with the current state of the world and works forwards, via an action that can be executed right away, to a successor state. This process is then repeated until a state is reached in which the desired goal is attained.

Could it be the case that non-human animals who solve planning problems do so in a way that resembles an AI planner using either backward chaining or forward chaining? If they do, then it is possible in principle to discriminate experimentally between the backward-chaining hypothesis and the forward-chaining hypothesis, using an extension of the experimental setup described by Bird & Emery (2009). Instead of stones and wide and narrow tubes, the extended apparatus uses multiple tubes, each with a cap with a uniquely shaped hole cut in it (Fig. S1). Only an object that matches the hole can be used in that particular tube. Using this apparatus, it's possible to set up a variety of planning problems, and, by varying the combinations of caps and objects, to make every new trial effectively a first trial. In particular, it is possible to devise a pair of planning problems that differ in difficulty for a forward chaining method but not for backward chaining, as well as a complementary pair of planning problem that differ in difficulty for a backward chaining method but not for a forward chaining method (Fig. S2).

The range of options explored in AI is not exhausted by the forward chaining / backward chaining opposition, however. Planning techniques have also been devised that combine forward and backward chaining. For example, the GraphPlan algorithm interleaves a forward-chaining graph construction phase with a backward-chaining graph search phase (Blum & Furst, 1997). Perhaps, when animals plan, their brains use a hybrid method of this sort. Indeed, if circumstances resembling the setup just described arise in the wild, in which the animal is cued with both an initial state (the sight of the tools) and a goal state (the sight of the food) before it acts, it seems unlikely that either a blind forward search (disregarding the goal state) or a blind backward search (disregarding the initial state) would be favoured by evolution. If an animal's brain uses a hybrid forward and backward method, we would not expect to find a marked difference in performance between the two paired conditions using the apparatus just described.

All of the AI methods cited so far work at the level of so-called "primitive" actions, that is to say, actions that are immediately executable (under the right conditions), such as picking up a twig or approaching a stone. In *hierarchical planning* (or hierarchical task network planning), by contrast, a suitable sequence of actions is found by refining a high-level action (such as "get food") into a sequence of lower-level sub-actions (such as "get a tool then retrieve the bucket"), then repeating this process until a plan comprising just primitive actions is reached (such as "get the wire, bend the wire, insert the wire into the tube, …"). Hierarchical organisation of behaviour is also implicit in many AI systems, such as those based on the cognitive architectures SOAR or ACT-R (Langley, *et al.*, 2009). Could there, perhaps, be an element of hierarchical decomposition in the way the brain of an animal arrives at a novel sequence of actions to attain a goal? If this were true, we would again expect no performance differential between the two paired conditions in the proposed sequential tool-use experiment.

3. Explaining Cognitive Prowess in Terms of Neurodynamics

The guiding metaphor in the kind of explanation envisaged above is that of classical, serial computation in the von Neumann style. To properly fill out such explanations, an account of how the proposed computational mechanisms are realised in the brain must be supplied. A variety of contemporary proposals seek to provide such an account (Duncan, 2010; Zylberberg, *et al.*, 2011). However, a feature common to all general-purpose planning methods is the need to carry out combinatorial search, and this is where their computational burden chiefly lies. The usual way to carry out combinatorial search on a conventional computer is with the aid of a dynamic data structure such as a stack. A *stack* is a structure that grows and shrinks over time in so-called "last-in-first-out" order, like a pile of plates only the top of which is accessible. In a search application, the stack is used to record branches of the tree of possibilities that remain to be explored. Although a plausible account can be constructed of how the brain might realise a form of serial processing resembling that of a conventional computer, it is less than obvious how such an account could convincingly be extended to describe the neural analogue of a stack.

Moreover, to implement a search algorithm on a serial virtual machine realised in a neural substrate would be to squander the brain's enormous potential for parallelism. Undoubtedly there are cognitive processes that are best carried out in a strictly serial order so that the effects of one operation are available for the next, such as mental arithmetic (Sackur & Dehaene, 2009) or the inner rehearsal of a sequence of actions (Shanahan, 2006). Indeed, for planning problems with minimal search requirements, a few serial steps of inner rehearsal to anticipate the outcome of a two- or three-action plan might be adequate. Perhaps the sequential tool use paradigms of Bird & Emery (2009) and Wimpenny, *et al.* (2009) fall into this category, because the challenges they involve can be met without departing from a narrow set of possibilities for action that are given by the learned affordances of the apparatus. In computer science terms, the branching factor of the tree of (obvious) possibilities is small, whether it is searched by backward chaining or forward chaining, and the reward can be obtained without moving beyond the obvious.

By contrast, what is impressive about Betty's performance in the wire-bending experiment of Weir, *et al.* (2002) is that it apparently combines experience from two micro-domains of expertise that were not previously associated, namely the domain of bending pliable materials and the domain of using tools to retrieve food. [4] That her brain was able to blend the relevant brain processes together is impressive enough. But even more striking is that her brain assembled the necessary raw ingredients for the blend in the first place, given no indication of their mutual relevance to the problem at hand (or at beak, rather). The branching factor of the tree of non-obvious possibilities for action is huge, which puts us firmly back in combinatorial territory. Surely this tree of possibilites is not searched by a slow, serial virtual machine realised on a massively parallel substrate. Could it be instead that the underlying massively parallel system can effect the search without sifting through an infeasibly large number of combinatorial possibilities one at a time? Perhaps, out of the maelstrom of competitive and co-operative local interactions between populations of neurons, a global network state corresponding to a suitable plan can emerge,

[4] Accomplishments of this sort evoke the *frame problem*, in the philosophers' sense of that term (Fodor, 1983; Dennett, 1984; Shanahan, 2009). This is the challenge of explaining how all of an agent's beliefs that are relevant to a problem or situation can be brought to bear without incurring an impractical cost to explicitly discriminate relevant from non-relevant beliefs. This is obviously closely related to the issue of cognitive integration. As a solution to the frame problem, the present paper builds on the proposal of (Shanahan & Baars, 2005).

“coalescing” out of the available dynamical elements rather than being the result of a sequence of discrete computational steps.

This is a somewhat fanciful sounding description, but it is reminiscent of a wide class of computations that can be characterised in terms of local interactions among parallel elements that collectively converge to a minimal state (an *attractor*). One of the best known examples is the Hopfield net. A *Hopfield net* is a network of artificial neurons with recurrent (feedback) connections which confer attractor dynamics on it (Hopfield, 1982; Amit, 1989). The network is trained to “remember” a number of patterns (its attractors). Given an input pattern, the trained network converges to a state corresponding to the nearest attractor, that is to say, to the remembered pattern that most closely resembles the input. It can function, therefore, as a form of associative memory. A Hopfield net, like any neural network, is inherently parallel, and each successive state is the product of strictly local computation. In other words, when a neuron is updated, its new state is a function only of the state of its neighbours, the neurons it is directly connected to.

Attractor networks have many practical applications (Smith, 1999), and have been proposed as fundamental building blocks of the brain (Amit, 1989; Freeman, 1999, Ch. 4; Rolls, 2009; Lansner, 2009). However, our interest here is not in building blocks, not in mere components of the brain. Rather, we are interested in how the brain’s global dynamics can settle on a pattern of activity that results in effective motor output given a novel challenge. To see how a global solution can arise from local processing with brain-scale parallelism, we need to look at networks with more complex topologies than a conventional attractor network. Moreover, the brain exhibits much richer dynamics than a conventional attractor network, including complex temporal patterns of spiking, metastability, rhythmic activity, and synchronisation phenomena. In the following sections, the challenge of achieving cognitive integration, as exemplified by Betty’s wire-bending, is recast as a problem of “coalition formation” in a dynamical system of this sort.

4. The Coalition Formation Problem

In computer science, we know how to exploit parallelism in various narrow domains. For example, the simulation of physical phenomena such as the diffusion of smoke in a burning building or the flow of air over a wing is most efficiently carried out using parallel computation. Likewise, in computer vision, basic operations like edge detection or optical flow are straightforwardly expressed in terms of parallel algorithms. In these cases, the breakdown of the overall problem into numerous sub-problems that can be carried out independently falls right out of the spatial structure of the problem domain. These so-called embarrassingly parallel problems are solved in the brain too. The visual cortex, for example, exploits neural parallelism in order to detect edges and motion rapidly. But this sort of parallelism is not much use for achieving cognitive integration. When an animal confronts a novel situation, the problem of drawing on the totality of its past experience to come up with a sequence of motor outputs that is effective despite never having been tried before is not one that can be straightfowardly broken down into trivial sub-problems.

To see the difficulty in its proper context, we need to think of the brain not so much as a set of parallel, distributed *computational* processes that take input, produce output, and transmit data to each other, but rather as a collection of parallel, distributed *dynamical* processes that unfold in continuous (rather than discrete) time and exert continuous influence on each other (Fig. 1). This is the character of the wetware that is biology's provision to cognition. Within this wetware substrate, the past experience of an animal is rendered into a set of brain processes that, both individually and in combination, are specialised for particular situations. The ongoing construction, maintenance and expansion of this set of specialists, and the increase in specialist expertise that results, is the responsibility of learning in the classical sense (Sutton & Barto, 1998; Shettleworth, 2010, Ch. 4). But specialist expertise is limited. A hallmark of cognitive sophistication is the capability to move beyond the tried-and-tested, to transcend domain-specific expertise, and thereby to tap the full combinatorial potential of an established sensorimotor repertoire. [5] The basic question of cognitive integration, then, is how the elements of an ever-changing, ever-expanding repertoire of specialist brain processes can be combined in previously untried coalitions.

The question, in other words, is how *open-ended* coalition formation is possible in a dynamical system like the brain. Implicit in the idea of coalition formation here is the idea of a *winning* coalition. One coalition has to arise that dominates the dynamics of the brain, shuts out all rivals, and dictates the animal's behaviour. A winning coalition will be in the ascendant only briefly. When events move on it will be

[5] We are again reminded of the frame problem here, as well as Mithen's (1996) notion of *cognitive fluidity*.

supplanted by a successor. But its victory, albeit temporary, must be complete. An animal whose brain allowed incompatible possibilities for action equal influence on its final motor output would not exhibit coherent behaviour. An animal cannot simultaneously move to the left and to the right, for example. It has just one body and this body is spatially confined and subject to various kinematic constraints. So the question of how open-ended coalition formation is possible is really the question of how *competitive* open-ended coalition formation is possible.

The basic question of open-ended coalition formation naturally leads to a number of subsidiary questions. In particular, how is it possible for a new coalition of sensory, motor, and memory processes to form that is *effective*, that results in behaviour that is more beneficial to the animal than if it had stuck to tried-and-tested combinations? There is no point in entertaining possibilities that are plain daft, for example. It does no good to an animal confronted with a food item that is just out of reach to initiate a courtship dance. But manipulating materials that resemble tools is promising. In short, candidate members of a successful coalition must be *relevant* to the situation at hand (or at beak). However, not every combination of relevant processes is effective. In variations of the trap-tube test (Seed, *et al*., 2006), pulling the plunger one way is *prima facie* as relevant as pulling it the other way. But one way results in the loss of the food reward, while the other way results in its attainment. The ideal system will freely explore novel coalitions of relevant processes, but will bias the competition between them according to their expected outcomes.

So the mechanism of coalition formation is not going to entertain every coalition with equal probability. The probability distribution in question will be shaped by experience. But our present interest is the connectivity and neurodynamics that makes open-ended coalition formation possible in the first place. How is it that a coalition of anatomically distributed brain processes can form whose constitution is not hardwired in advance? Cognitive integration is achieved when the animal is able to draw upon the full battery of its neuronal resources, mixing and matching them as required, to find an effective, and sometimes innovative, response to unfamiliar situations. A winning coalition, a coalition of brain processes that issues in overt behaviour, will comprise just a small subset of those resources. But to the extent that cognitive integration is achieved, the whole brain will have participated in an open competition to find that subset.

5. Brain Connectivity

When one computer exchanges information with another through the Internet, the information in question is chopped up into packets, and each packet is sent separately to its destination. Packets can arrive in a different order from that in which they were sent (to be re-assembled in the correct order on arrival), and different packets can go via different routes. So the physical connections that underlie this process give away very little about the flow of information in the network they support. But the brain of an animal works very differently. If one brain region is to exercise a direct influence on another, it must be directly connected to that region. So the pattern of physical connections between brain regions is good evidence of how information and influence flow around the brain. It makes sense, therefore, to study brain connectivity with a view to understanding how arbitrarily constituted coalitions of anatomically distributed brain processes form.

The large-scale network organisation of an animal's brain is revealed in several steps (Bullmore & Sporns, 2009; Sporns, 2010). First, anatomical investigations yield data about the existence of pathways between brain regions, either using traditional tracing methods or using diffusion-based imaging *in vivo*. This data is then compiled into a connectivity matrix covering a major sub-division of the animal's forebrain — the cerebral cortex, say, or perhaps the entire telencephalon. This is the animal's (structural) *connectome*. The resulting connectivity matrix is is then subjected to analysis using the mathematical theory of complex networks (or "graphs"), which reveals its large-scale organisation. The animal that has been most thoroughly investigated in this way is the human (Hagmann, *et al.*, 2008; Iturria-Medina, *et al.*, 2008; Gong, *et al.*, 2009, van den Heuvel & Sporns, 2011). But connectivity matrices and network analyses have been carried out for a number of other species, including the fruit-fly (Chiang, *et al.*, 2011), the cat (Scannell, *et al.*, 1999; Sporns, *et al.*, 2007; Zamora-López, *et al.*, 2010), the macaque (Sporns, *et al.*, 2007; Modha & Singh, 2010), and the pigeon. [6]

A number of topological features consistently arise in these studies. These include small-world organisation and modularity. A sparse network (all brain networks are sparse) is a *small world* if it has a mean path length comparable to, but a

[6] The pigeon connectome is the subject of ongoing collaborative work by the present author, Vern Bingman, Toru Shimizu, Martin Wild, and Onu Güntürkün.

clustering coefficient higher than, an equivalent randomly wired network (Watts & Strogatz, 1998) (Fig. 2A). A network's *mean path length* is the average length of the shortest path between any two nodes, while its *clustering coefficient* is the proportion of a node's neighbour's that neighbour each other averaged over the whole network. A small world network retains a capacity for locally concentrated processing, but facilitates the rapid flow of information around the network as a whole. A network is *modular* if it can be partitioned into subsets of nodes (modules) that are densely connected internally but only sparsely connected to other subsets (Girvan & Newman, 2002) (Figs. 2B). Modules are suggestive of functional specialisation.

Another consistent feature of brain networks is the presence of topologically significant *hub nodes* (Sporns, *et al.*, 2007). Although different authors use slightly different criteria for designating a node as a hub, the key attribute in all cases is high connectivity. A hub node is one that has an unusually large number of connections, or one that lies on an unusually large number of short paths. Hub nodes can be further classified in the context of a network's modularity. A *connector hub* is one that plays an important role in communication between modules (Fig. 2C), while a *provincial hub* is one that plays an important role in communication within a module. The human, macaque, and cat connectomes have all been shown to possess a pronounced *connective core* of topologically central hub nodes (Hagmann, *et al.*, 2008; Zamora-López, *et al.*, 2010; Modha & Singh 2010; van den Heuvel & Sporns, 2011).

The addition of the pigeon to this list would be significant because it serves as a prototypical avian species, and it would be surprising indeed if the same network properties were not present in the (larger) forebrains of corvids. We might expect the set of hub nodes that comprises the connective core in the pigeon to be constituted by brain regions that are either homologous to, or functionally analogous to, connective core regions in humans and macaques, such as the hippocampal formation, which is implicated in long-term memory, and the nidopallium caudolaterale, a prefrontal-like region which is implicated in goal-directed action and working memory. The presence in all these species of a connective core that includes a prefrontal functional analogue as well as homologous hippocampal structures would support the hypothesis of convergent evolution of intelligence in apes and birds (Emery & Clayton, 2005; Seed, *et al.*, 2009), suggesting a common neural substrate for cognition.

6. The Connective Core Hypothesis

What are the specific cognitive implications of the fact that an animal's forebrain has the network properties just listed? The contention of this paper is that a brain's capacity for cognitive integration depends on the presence within the telencephalon of a pronounced connective core, that is to say a small set of hub nodes that are topologically central to the whole network and richly connected to each other (Shanahan, 2010a, Ch. 4). Its topologically central situation entails that information and influence can funnel in to and fan out from the connective core, which makes it ideally suited to fulfil three functional roles that together promote the availability of the brain's full repertoire of process combinations. According to the hypothesis, the connective core is

1) a locus of broadcast,

2) a medium of coupling, and

3) an arena for competition.

The connective core is a potential locus of broadcast because the ongoing pattern of activity in the regions comprising it can exercise direct influence on a large number of other forebrain regions, and indirect influence on an even larger number via just a few hops in the network. Suppose an animal is presented with an unexpected object (such as a straight piece of wire in a tool manufacture paradigm). The resulting visual stimulus excites its visual areas, which in turn can influence activity in its connective core. Thanks to the core's rich connections to other regions, this influence can then be disseminated throughout the forebrain, arousing the widespread activation of multiple neuronal groups (Baars, 1988; Dehaene, *et al.*, 2006), including those associated with the visible affordances of the object (it can be pushed, picked up, pecked at, used as a prod, and so on).

To the extent that the connective core fulfils this broadcast role, it addresses one aspect of the coalition formation problem, namely how the set of brain processes relevant to the ongoing situation becomes active and is thereby given the opportunity to join a coalition that might eventually take over the animal's motor apparatus, while those that are irrelevant are excluded (Shanahan & Baars, 2005). The relevance or otherwise of a process is not decided centrally, in any sense. Thanks to the connective core, this responsibility is delegated to the processes themselves, which carry it out individually and in parallel. But cutting down the set of active processes to only those that are relevant is only half the problem. Even the pool of relevant brain processes may be large, and the set of possible coalitions constituted by members of that pool is combintaorially larger.

One consequence of the combinatorics is the prohibitive quantity of wiring that would be needed if each group of neurons was directly connected to every other group of neurons with which it might enter into a coalition. This is where the second purported role of the connective core comes into play. In much the same way that activity within the connective core can influence any part of the forebrain, so any part of the forebrain can influence the connective core and, via the connective core, any other part of the forebrain. So the connective core has the potential to act as a communications infrastructure capable of routing information and influence between any two brain processes, allowing them to become coupled (Shanahan, 2010a, Ch. 4 & 5; Zylberberg, *et al*., 2010). Thanks to its covergent-divergent wiring pattern, the connective core allows arbitrary combinations of processes to become coupled without recourse to combinatorial wiring. This addresses the most prominent aspect of the coalition formation problem from the standpoint of cognitive integration, as it allows novel coalitions to arise.

However, like any channel of communication, the connective core has only limited bandwidth. Rival coalitions must compete for this bandwidth. In a telephone network, contention for limited bandwidth is resolved equitably. In the early days of telephony, customers on a busy trunk line had to wait until a channel became available, while nowadays they might have to tolerate a reduction in their quality of service. Either way, everyone gets a fair share of the bandwidth. By contrast, contention for access to the connective core is not resolved equitably, according to the present hypothesis. Rather, it is resolved through winner-takes-all competition. The coalition that ends up driving activity in the connective core will shut out its competitors by preventing their constituent processes from exchanging influence and information through it. In this sense, the competition plays out in the connective core itself. It thereby addresses a further aspect of the coalition formation problem, namely the need for a single coalition at a time to dominate the dynamics of the brain and dictate the animal's behaviour.[7]

So far we have been imagining a single transition in the state of the brain. We have exmanined the putative role played by the connective core in producing the globally integrated brain state that determines an animal's response to its current situation. So the emphasis up to now has been on parallel processing. But the distinctive blend of serial and parallel processing supported by the connective core involves many such transitions. Insofar as the connective core has only a limited capacity, it constitutes a *central bottleneck* in the brain (Pashler, 1984; Marois & Ivanoff, 2005). Apart from helping to enforce the sort of winner-takes-all competition that is necessary for an animal to commit to a coherent course of action, this central bottleneck facilitates serial processing, which is necessary for the chaining of mental operations (Sackur & Dehaene, 2009). Thanks to its limited capacity, a serial procession of states emerges via the connective core. Yet each state-to-state transition in this series is the product of competition and co-operation among massively numerous parallel processes (Fig. 3).

This unconventional model of computation can be characterised in dynamical systems terms. The brain settles into global, attractor-like states, mediated by the connective core. But these states, though attractor-like, are only temporary. That is to say, they are not stable but *metastable* (Kelso, 1995; Bressler & Kelso, 2001; Werner, 2007; Deco, *et al*., 2011). A coalition is held together by its own complex dynamics, which will eventually precipitate its break-up. (A likely feature of this complex dynamics is synchronous oscillation (Fries, 2009; Shanahan, 2010b; Shanahan & Wildie, 2012).) Moreover, in a behaving animal, the brain is subject to external perturbation due to incoming sensory stimulation, which can similarly upset the stability of a coalition. So the overall brain-wide pattern of activity is one of coalition formation, followed by break-up, followed by the formation of a new coalition, and so on, yielding a serial procession of global metastable states punctuated by transients.

[7] Competitive mechanisms have long been thought to play an important part in brain dynamics (Desimone & Duncan, 1985). The need for the brain to settle into an integrated global state that determines a coherent behavioural outcome when faced with competing opportunities or threats has been dubbed the *action selection problem* (Prescott, 2007). Various structures, notably the basal ganglia, have been hypothesised to address the action selection problem (Redgrave, *et al*., 1999). The same competitive mechanisms are presumed to be implicated in the competition for the connective core.

One way to think of the blend of sequential and parallel processing envisioned here is in terms of a serial *virtual machine* realised on a massively parallel neural substrate (Dennett, 1991, pp.209–226; Sloman & Chrisley, 2005). However, the processing carried out by this serial virtual machine is unlike that of a conventional von Neumann computer. In the present case as in a conventional computer we have states, but those states possess an ongoing internal dynamics that is absent in conventional computation. In the present case as in a conventional computer we have transitions from one state to another, but these transitions are here mediated by a rich temporal dynamics that is quite different from the discrete state-to-state transitions found in conventional computation.

Inherent in the way the present system operates is the possibility of dealing with a cognitive challenge either through a series of sequential internal steps, as in a conventional AI planning algorithm, or through a form of parallel search resembling the convergence of an attractor network, or indeed through some combination of both methods. However, continuing the comparison with a von Neumann architecture, we might ask what, in the envisaged serial virtual machine, is the analogue of memory in a conventional computer. For the serial machine to carry out anything resembling the sequential processing of an AI problem solving algorithm, it must be possible for certain aspects of the present state to endure and influence future states. This is accomplished in the framework of this paper in the following way. When the presently dominant coalition of processes breaks up and gives way to a successor, certain members of the outgoing coalition can remain active, retaining information, possibly to join a later dominant coalition. There is ample evidence — from lesion studies with delayed response tasks, for example — that mnemonic brain processes of this sort are supported by the prefrontal cortex in mammals (a member of the connective core), and by its analogue in birds (the nidopallium caudolaterale) (Fuster, 2008; Güntürkün, 2005).

To summarise, the presence of a small set of topologically central forebrain regions has been established empirically. The potential implications of this finding for understanding animal cognition, and perhaps for understanding animal consciousness, are profound. They are encapsulated by the following slogan: unity from multiplicity, serial from parallel. Unity arises from multiplicity because the connective core ensures that a singular, coherent response to the ongoing situation is integrated from the brain’s full resources. Serial emerges from parallel because brain-wide information and influence are channeled through the connective core, which acts as a limited bandwidth processing bottleneck, allowing mental operations to be chained together.

7. Synthetic Methodology

The connective core hypothesis purports to explain how the brain's connectivity and dynamics enable certain kinds of sophisticated cogntion that have been demonstrated experimentally in non-human animals. How can we test this hypothesis and, insofar as it is true, build a profile of the serial and parallel activity involved in the successful performance of different tasks? The behavioural paradigm described earlier is one approach. But to really get at the underlying mechanisms, such behavioural methods need to be supplemented. Although neuroanatomy, imaging, and electrophysiological recording are vital sources of relevant data, the focus of this final section is a complementary but less well-established methodology, namely the construction of computer and robot models.

As Braitenberg (1984) noted, in his law of "uphill analysis and downhill invention", it is often more difficult to infer principles of internal operation from experimental observation than it is to invent mechanisms that produce the same outward behaviour. Invention cannot be a substitute for observation, because more than one mechanism is typically capable of generating a given behaviour. What we seek are computer and robot models that not only generate the required behaviour, but are also consistent with the empirical findings of neuroscience. A computer program is not a scientific theory, of course, and nor is a working robot. But properly engineered computer programs and robots are built according to principles and specifications, and it is these principles and specifications, rather than any artefact conforming to them, that contribute to our scientific understanding. A computational model of a neuroscientific hypothesis is a demonstration that the hypothesis is sufficiently clear to be implementable. Moreover, the rigours of implementation often reveal unexpected nuances of a hypothesis, as well as extensions and alternatives that might otherwise have been missed.

This is the rationale behind the field of computational neuroscience. But here we are advocating the extension of this synthetic methodology to the embodied case, to the use of robots. This approach has been applied to various brain structures, including the hippocampus (Fleischer, *et al*., 2007), the basal ganglia (Prescott, *et al.*

2006), and the amygdala (Ziemke & Lowe, 2009), as well as to low-level behaviours such as phonotaxis in invertebrates (Webb, 2002). (For an overview of recent work, see Krichmar & Wagatsuma (2011).) In lieu of detailed experimental data, can the synthetic approach be used to bolster the connective core hypothesis, and the claim that it can explain aspects of animal cognition?

Some steps in this direction have already been taken. In Shanahan (2006), a cognitive architecture is presented whose key component is a network of artificial neurons (the *global workspace* (Baars, 1988)) that fulfils the same role as that imputed to the connective core. It is an arena for competition among rival groups of neurons, each of which attempts to make it settle into a particular pattern of reverberating activation. The winning pattern is broadcast back to the full cohert of competitors, giving rise to a cycle of competition followed by broadcast. Thus, a serial procession of states is produced, but each serial step is the outcome of parallel processing in tens of thousands of artificial neurons. This serial process is used to rehearse the consequences of actions before they are carried out. The resulting predictions modulate the salience of possible actions in the context of a winner-takes-all action selection mechanism, which controls a simple simulated robot. The resulting model was used to emulate a classic experiment of Tolman and Gleitman (1949) that appeared to show a capacity for look-ahead in rats (Fig. S3).

The work just described showed that a global neuronal workspace — a structure analogous to a connective core — could form the basis of serial computation in a massively parallel setting, and demonstrated its deployment in a control architecture for a robot. However, the model used a form of artificial neuron that lacks biological fidelity. A more realistic computer model of competitive access to a global neuronal workspace was developed by Deheane and colleagues using spiking neurons (Dehaene, *et al.*, 2003; Dehaene & Changeux, 2005). A complementary spiking model of global neuronal workspace broadcast was later reported in Shanahan (2008) (Fig. 4), and then elaborated by Connor and Shanahan (2010). A related model showcasing the possible role of the brain's connective core in inter-process communication was described by Zylberberg, *et al.* (2010).

While these models contain many of the elements required to instantiate the connective core hypothesis, they are also lacking certain crucial features. In particular, they do not address the issue of coalition formation, which is central to the present paper. On the assumption that elevated levels of synchronous oscillation in spatially separated brain regions are a sign that those regions are operating as a coalition, the model described Shanahan (2010b) is a step in the right direction. The model is based on oscillators, and is thus defined at a higher level than spiking neurons. Moreover, although its connectivity is modular and small-world, it lacks a connective core. Nevertheless, the model shows how metastable coalitions can form and break up in a modular, small-world network of dynamical elements. Cabral, *et al.* (2011) have shown that a similar setup can be used to model human resting state fMRI data, including the default mode network in which the regions of the human connective core are prominent.

A more complete implementation of the connective core hypothesis, one that could actually exhibit cognitive integration, would combine the important features of each of the models described. The first step would be to reverse engineer the basic blueprint of the vertebrate brain, something akin to what Striedter (2005) calls the "vertebrate brain archetype" (after the 19th Century biologist Richard Owen). This should describe the simplest viable brain that incorporates every structure common to all vertebrates (whether as homologues or functional analogues), and retain those connectivity properties that are similarly shared. The next step would be to construct a model brain according to this blueprint that conforms to the connective core hypothesis.

The key features of this model would be 1) biologically realistic network topology, specifically a hierarchically modular small-world network with a connective core of hub nodes, 2) functional differentiation to include at least sensory, motor, basal gangliar, amygdaloid, hippocampal, and prefrontal regions, and 3) embodiment in a behaving robot endowed with a basic motivational system and associative learning mechanisms. To be consistent with the framework of the present paper, the connective core would be expected to exhibit certain dynamical properties, namely 1) episodes of (metastable) broadcast punctuated by competition, and 2) the formation of coalitions of brain processes drawn from a large, open-ended repertoire. If the proposals of the present paper are correct then, by scaling up the model in different ways, it ought be possible to replicate the cognitive feats of a range of species, including those that seemingly demonstrate "insight".

8. Concluding Remarks

There is growing evidence that the brains of cognitively sophisticated animals are hierarchically modular small-world networks with connective cores comprising sets of connector hubs. According to the hypothesis of this paper, possession of a connective core is a prerequisite for sophisticated animal cognition. As well as unifying the distributed activity of the brain's massively parallel resources, the connective core promotes cognitive integration by allowing the formation of novel coalitions of brain processes, and facilitates serial mental operations. However, possession of a connective core is no guarantee of cognitive sophistication. Indeed, the evolutionary pressures that led to this network organisation may at first have been less to do with cognitive prowess than to do with the space and energy costs of connecting large numbers of neurons to each other.

It is typically taken for granted that the most parsimonious explanation of an animal's behaviour is to be preferred in the absence of evidence ruling it out. Yet it is not always obvious what constitutes pasimony (Heyes, 2012). A simple explanation of a simple behaviour may look like the right choice at the level of a single species. But from a wider scientific perspective, a unified theory that encompassed the full spectrum of animal behaviour would be preferable. The goal of supplying such a theory is likely to be furthered by a better understanding of the underlying neural mechanisms. It may be the case, for example, that behaviours that can be given an associative account and behaviours that seem to demand a cognitive account both emerge from the same underlying network organisation. Perhaps the neural underpinnings of "insight", far from being mysterious, are prevalent in large centralised nervous systems, thanks to massive parallelism, rich dynamics, and a particular network organisation.

Acknowledgments

Thanks to Jackie Chappell, Lucy Cheke, Nicky Clayton, Uri Grodzinski, Onur Güntürkün, Richard Newcombe, and Alex Taylor.

References

Amit, D.J. (1989). *Modeling Brain Function: The World of Attractor Neural Networks*. Cambridge University Press.

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

Beck, S.R., Apperly, I.A., Chappell, J., Guthrie, C. & Cutting, N. (2011). Making Tools Isn't Child's Play. *Cognition* 119, 301–306.

Bird, C.D. & Emery, N.J. (2009). Insightful Problem Solving and Creative Tool Modification by Captive Nontool-using Rooks. *Proceedings of the National Academy of Science USA* 106 (25), 10370–10375.

Blum, A.L. & Furst, M.L. (1997). Fast Planning Through Planning Graph Analysis. *Artificial Intelligence* 90 (1–2), 281–300.

Braitenberg, V. (1984). *Vehicles: Experiments in Synthetic Psychology*. MIT Press.

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

Bullmore, E. & Sporns, O. (2009). Complex Brain Networks: Graph Theoretical Analysis of Structural and Functional Systems. *Nature Reviews Neurosci*ence 10, 186–198.

Cabral, J., Hugues, E., Sporns, O. & Deco, G. (2011). Role of Local Network Oscillations in Resting-State Functional Connectivity. *NeuroImage* 57, 130–139.

Chiang, A.-S. *et al.* (2011). Three-Dimensional Reconstruction of Brain-wide Wiring Networks in Drosophila at Single-Cell Resolution. *Current Biology* 21, 1–11.

Chittka, L. & Niven, J. (2009). Are Bigger Brains Better? *Current Biology* 19, R995–R1008.

Connor, D. & Shanahan, M.P. (2010). A Computational Model of a Global Neuronal Workspace with Stochastic Connections. *Neural Networks* 23, 1139–1154.

Deco, G., Jirsa, V.K. & McIntosh, A.R. (2011). Emerging Concepts for the Dynamical Organization of Resting-State Activity in the Brain. *Nature Reviews Neuroscience* 12, 43–56.

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

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J. & Sergent, C. (2006). Conscious, Preconscious, and Subliminal Processing: A Testable Taxonomy. *Trends in Cognitive Sciences* 10 (5), 204–211.

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 USA* 100 (14), 8520–8525.

Dennett, D. (1984). Cognitive Wheels: The Frame Problem in Artificial Intelligence. In C.Hookway (ed.), *Minds, Machines and Evolution*, Cambridge University Press, pp. 129–151.

Dennett, D. (1991). *Consciousness Explained*. Penguin Press.

Duncan, J. (2010). The Multiple-Demand (MD) System of the Primate Brain: Mental Programs for Intelligent Behaviour. *Trends in Cognitive Sciences* 14 (4), 172–179.

Epstein, R. (1985). Animal Cognition as the Praxist Views It. *Neuroscience and Biobehavioural Reviews* 9, 623–630.

Emery, N.J. & Clayton, N.S. (2005). Evolution of the Avian Brain and Intelligence. *Current Biology* 15 (23), R946–R950.

Fleischer, J.G., Gally, J.A., Edelman, G.M. & Krichmar, J.L. (2007). Retrospective and Prospective Responses Arising in a Modeled Hippocampus During Maze Navigation by a Brain-based Device. *Proc. Nat. Acad. Sci. USA* 104 (9), 3556–3561.

Fodor, J.A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. MIT Press.

Freeman, W.J. (1999). *How Brains Make Up Their Minds*. Weidenfeld & Nicolson.

Fries, P. (2009). Neuronal Gamma-Band Synchronization as a Fundamental Process in Cortical Communication. *Annual Review of Neuroscience* 32, 209–224.

Fuster, J. (2008). *The Prefrontal Cortex (Fourth Edition)*. Academic Press.

Girvan, M. & Newman, M.E.J. (2002). Community Structre in Social and Biological Networks. *Proceedings of the National Academy of Science* USA 99, 7821–7826.

Gong, G., He, Y., Concha, L., Lebel, C., Gross, D.W., Evans, A.C. & Beaulieu, C. (2009). Mapping Anatomical Connectivity Patterns of Human Cerebral Cortex Using In Vivo Diffusion Tensor Imaging Tractography. *Cerebral Cortex* 19, 524–536.

Güntürkün, O. (2005). The Avian 'Prefrontal Cortex' and Cognition. *Current Opinion in Neurobiology* 15, 686–693.

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C.J., Wedeen, C.J. & Sporns, O. (2008). Mapping the Structural Core of Human Cerebral Cortex. *PLoS Biology* 6 (7), e159.

Hansell, M. (2007). *Built by Animals: The Natural History of Animal Architecture*. Oxford University Press.

Heyes, C. (2012). Simple Minds: A Qualified Defence of Associative Learning. This volume.

Hopfield, J. (1982). Hopfield, J. J. (1982). Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings of the National Academy of Science USA* 79, 2554–2558.

Iturria-Medina, Y., Sotero, R.C., Canales-Rodríguez, E.J., Alemán-Gómez, Y. & Melie-García, L. (2008). Studying the Human Brain Anatomical Network via Diffusion-weighted MRI and Graph Theory. *NeuroImage* 40, 1064–1076.

Kelso, J.A.S. (1995). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. MIT Press.

Krichmar, J.L. & Wagatsuma, H. (2011). *Neuromorphic and Brain-based Robots*. Cambridge University Press.

Langley, P., Laird, J.E. & Rogers, S. (2009). Cognitive Architectures: Research Issues and Challenges. *Cognitive Systems Research* 10, 141–160.

Lansner, A. (2009). Associative Memory Models: From the Cell-Assembly Theory to Biophysically Detailed Cortex Simulations. *Trends in Neurosciences* 32 (3), 178–186.

Marois, R. & Ivanoff, J. (2005). Capacity Limits of Information Processing in the Brain. *Trends in Cognitive Sciences* 9 (6), 296–305.

Mithen, S. (1996). *The Prehistory of the Mind*. Thames & Hudson.

Modha, D.S. & Singh, R. (2010). Network Architecture of the Long-distance Pathways in the Macaque Brain. *Proceedings of the National Academy of Science USA* 107 (30), 13485–13490.

Pashler, H. (1984). Processing Stages in Overlapping Tasks: Evidence for a Central Bottleneck. *Journal of Experimental Psychology: Human Perception and Performance* 10 (3), 358–377.

Prescott, T.J. (2007). Forced Moves or Good Tricks in Design Space? Landmarks in the Evolution of Neural Mechanisms for Action Selection. *Adaptive Behavior* 15 (1), 9–31.

Prescott, T.J., Montes González, F.M., Gurney, K. Humphries, M.D. & Redgrave, P. (2006). A Robot Model of the Basal Ganglia: Behavior and Intrinsic Processing. *Neural Networks* 19, 31–61.

Raby, C.R., Alexis, D.M., Dickinson, A. & Clayton, N.S. (2007). Planning for the Future by Western Scrub-Jays. *Nature* 445, 919–921.

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

Rolls, E.T. (2009). Attractor Networks. *Wiley Interdisciplinary Reviews: Cognitive Science* 1 (1), 119–134.

Russell, S. & Norvig, P. (2010). *Artificial Intelligence: A Modern Approach (Third Edition)*. Pearson.

Sackur, J. & Dehaene, S. (2009). The Cognitive Architecture for Chaining of Two Mental Operations. *Cognition* 111, 187–211.

Scannell, J.W., Burns, G.A.P.C, Hilgetag, G.C., O'Neil, M.A. & Young, M.P. (1999). The Connectional Organization of the Cortico-thalamic System of the Cat. *Cerebral Cortex* 9, 277–299.

Seed, A.M. & Byrne, R. (2011). Animal Tool-Use. *Current Biology* 20, R1032–R1039.

Seed, A., Emery, N. & Clayton, N. (2009). Intelligence in Corvids and Apes: a Case of Convergent Evolution? *Ethology* 115, 401–420.

Seed, A.M., Tebbich, S., Emery, N.J. & Clayton, N.S. (2006). Investigating Physical Cognition in Rooks, *Corvus frugilegus*. *Current Biology 16*, 697–701.

Shanahan, M.P. (2006). A Cognitive Architecture that Combines Internal Simulation with a Global Workspace. *Consciousness and Cognition* 19, 433–449.

Shanahan, M.P. (2008). A Spiking Neuron Model of Cortical Broadcast and Competition. *Consciousness and Cognition* 17, 288–303.

Shanahan, M.P. (2009). The Frame Problem. In *The Stanford Encyclopedia of Philosophy (Winter 2009 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2009/entries/frame-problem/>.

Shanahan, M.P. (2010a). *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds*. Oxford University Press.

Shanahan, M.P. (2010b). Metastable Chimera States in Community-structured Oscillator Networks. *Chaos* 20, 013108.

Shanahan, M.P. & Baars, B. (2005). Applying Global Workspace Theory to the Frame Problem. *Cognition* 98 (2), 157–176.

Shanahan, M.P., Bingman, V.P., Shimizu, T., Wild, M. & Güntürkün, O. (2012). Large-Scale Network Organisation in the Avian Forebrain. Submitted.

Shanahan, M.P. & Wildie, M. (2012). Establishing Communication Between Neuronal Populations Through Competitive Entrainment. *Frontiers in Computational Neuroscience* 5, Article 62.

Shettleworth, S.J. (2010). *Cognition, Evolution, and Behavior (Second Edition)*. Oxford University Press.

Sloman, A. & Chrisley, R.L. (2005). More Things than Are Dreamt of in Your Biology: Information-Processing in Biologically Inspired Robots. *Cognitive Systems Research* 6, 154–174.

Smith, K. (1999). Neural Networks for Combinatorial Optimization: A Review of More Than a Decade of Research. INFORMS Journal on Computing 11 (1), 15–34.

Sporns, O. (2010). *Networks of the Brain*. MIT Press.

Sporns, O., Honey, C.J. & Kötter, R. (2007). Identification and Classification of Hubs in Brain Networks. *PLoS One* 10, e1049.

Striedter, G.F. (2005). *Principles of Brain Evolution*. Sinaur Associates.

Sutton, R.S. & Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Taylor, A.H., Eliffe, D., Hunt, G.R. & Gray, R.D. (2010). Complex Cognition and Behavioural Innovation in New Caledonian Crows. *Proceedings of the Royal Society B* 277, 2637–2643.

Tolman, E. C., & Gleitman, H. (1949). Studies in Learning and Motivation: I. Equal Reinforcements in Both End-Boxes, Followed by Shock in One End-Box. *Journal of Experimental Psychology* 39, 810–819.

van den Heuvel, M.P. & Sporns, O. (2011). Rich-Club Organization of the Human Connectome. *Journal of Neuroscience* 31 (44), 15775–15786.

Watts, D.J. & Strogatz, S.H. (1998). Collective Dynamic of ‘Small-world’ Networks. *Nature* 393, 440–442.

Webb, B. (2002). Robots in Invertebrate Neuroscience. *Nature* 417, 359–363.

Weir, A.A.S., Chappell, J. & Kacelnik, A. (2002). Shaping of Hooks in New Caledonian Crows. *Science* 297, 981.

Werner, G. (2007). Metastability, Criticality, and Phase Transitions in Brain and its Models. *BioSystems* 90, 496–508.

Wimpenny, J. H., Weir, A.A.S, Clayton, L., Rutz, C. & Kacelnik, A. (2009). Cognitive Processes Associated with Sequential Tool Use in New Caledonian Crows. *PloS One* 4 (8), e6471.

Zamora-López, G., Zhou, C. & Kurths, J. (2010). Cortical Hubs Form a Module for Multisensory Integration on Top of the Hierarchy of Cortical Networks. *Frontiers in Neuroinformatics* 4, Article 1.

Ziemke, T. & Lowe, R. (2009). On the Role of Emotion in Embodied Cognitive Architectures: From Organisms to Robots. *Cognitive Computation* 1, 104–117.

Zylberberg, A. Dehaene, S., Roelfsema, P.R. & Sigman, M. (2011). The Human Turing Machine: a Neural Framework for Mental Programs. *Trends in Cognitive Sciences* 15 (7), 293–300.

Zylberberg, A., Fernández Slezak, D., Roelfsema, P.R., Dehaene, S. & Sigman, M. (2010). The Brain's Router: A Cortical Network Model of Serial Processing in the Primate Brain. *PloS Computational Biology* 6 (4), e1000765.

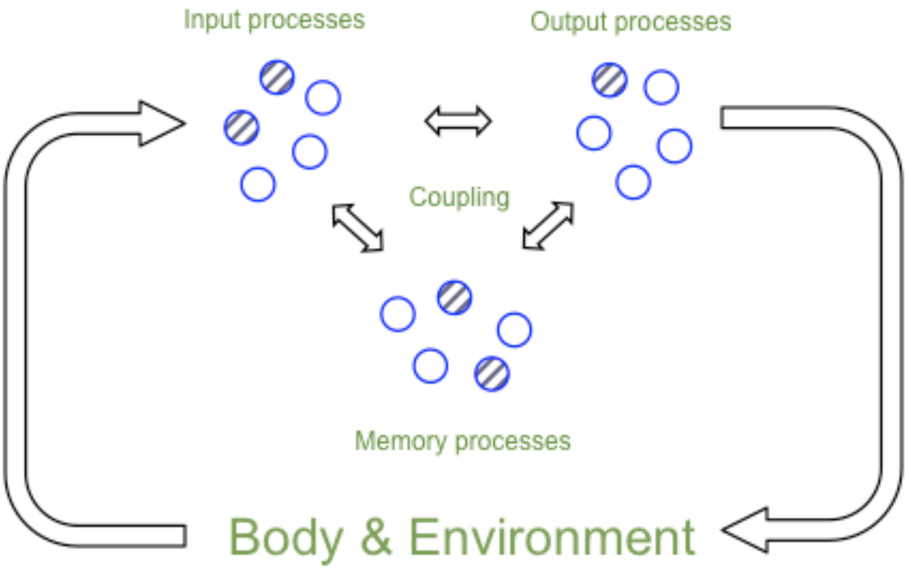

Figure 1: Competitive coalition formation. A coalition of coupled brain processes forms (shown as hatched circles) to the exclusion of rival coalitions, partly under the influence of internal dynamics (including the preceding coalition) and partly under the influence of external stimuli. Coalition membership is drawn from a large pool that includes sensory, motor, and memory processes. Brain processes are anatomically localised, functionally specialised populations of neurons, but a coalition of brain processes can be anatomically distributed and functionally diverse. The dominant coalition governs the behaviour of the animal until it breaks up and is succeded by a new coalition.

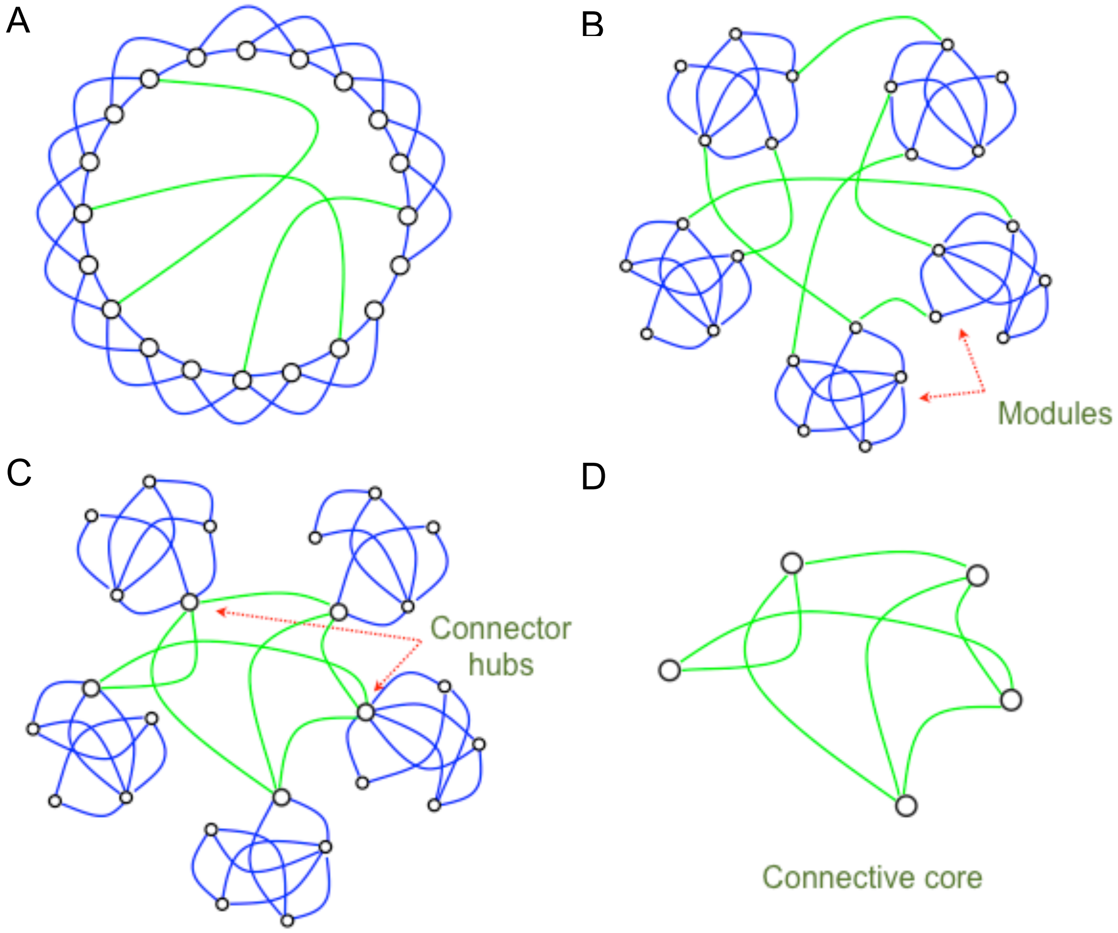

Figure 2: (A) A small-world network built using the Watts-Strogatz procedure (Watts & Strogatz, 1998). This network lacks modularity. (B) A modular small-world network that more closely resembles the large-scale structural connectivity of an animal's brain (but lacks connector hubs). (C) A modular small-world network with connector hubs. A node is designated a connector hub if it plays a significant role in communication between modules. (D) The connector hubs, along with their dense interconnections, are designated the connective core of the network.

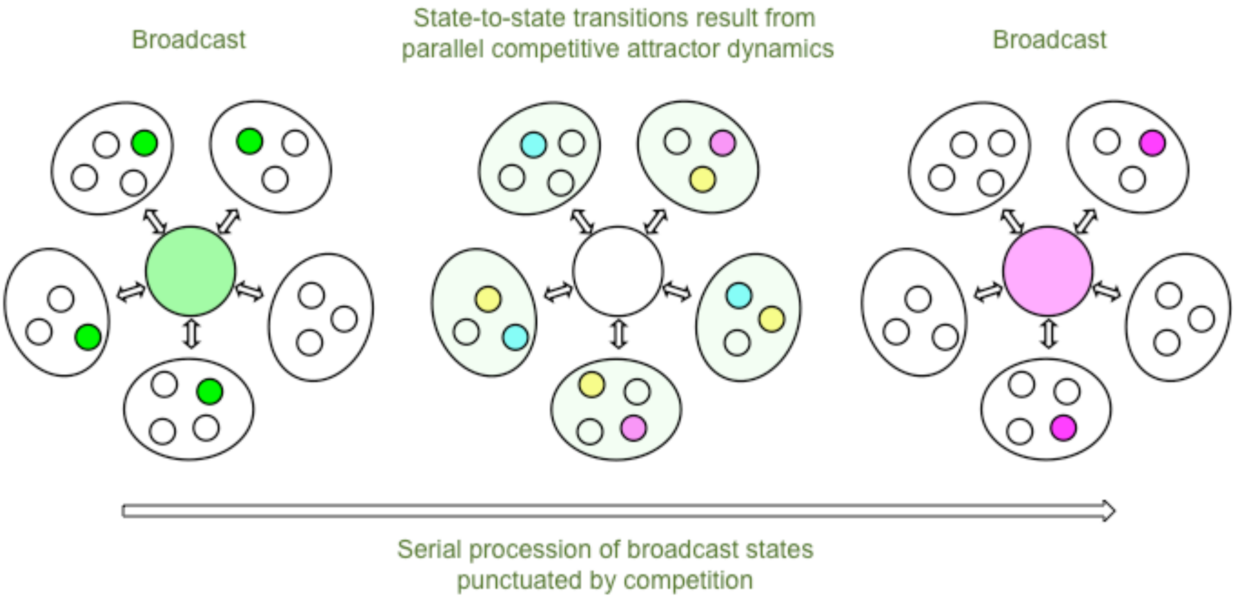
Broadcast
State-to-state transitions result from parallel competitive attractor dynamics
Broadcast
Serial procession of broadcast states punctuated by competition

Figure 3: The connective core supports a distinctive blend of serial and parallel processing. The transition from one broadcast state to another results from competition (differently coloured circles) and co-operation (similarly coloured circles) among parallel brain processes (here organised into five modules). In this case, the broadcast influence of the green coalition gives rise to a competition among three potential successors (yellow, cyan, and magenta). The magenta coalition triumphs, giving rise to the next broadcast state.

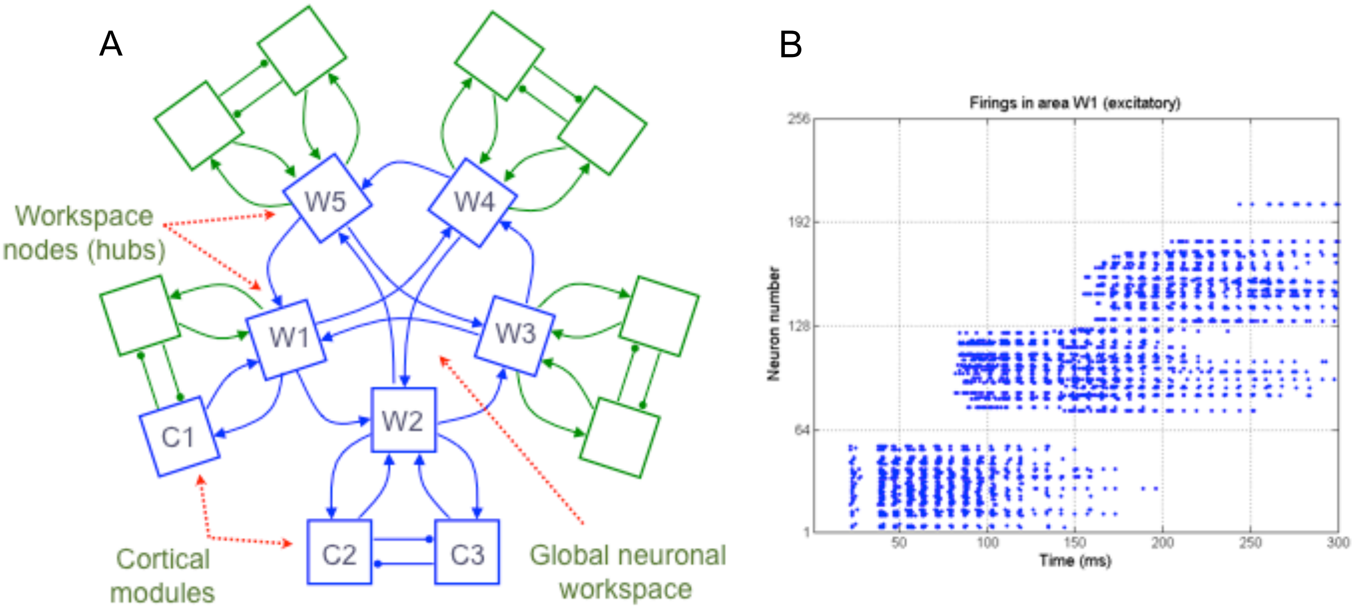

Figure 4: (A) The architecture of the spiking neuron model of a global neuronal workspace presented in Shanahan (2008). The workspace nodes and their interconnections are functional analogues of the connective core, according to the hypothesis of the present paper. (Compare Fig. 2C.) The blue portions were included in the computer simulation. (B) A raster plot of the output from the model, showing the sequential production of reverberating spatial patterns in the connective core (global workspace).

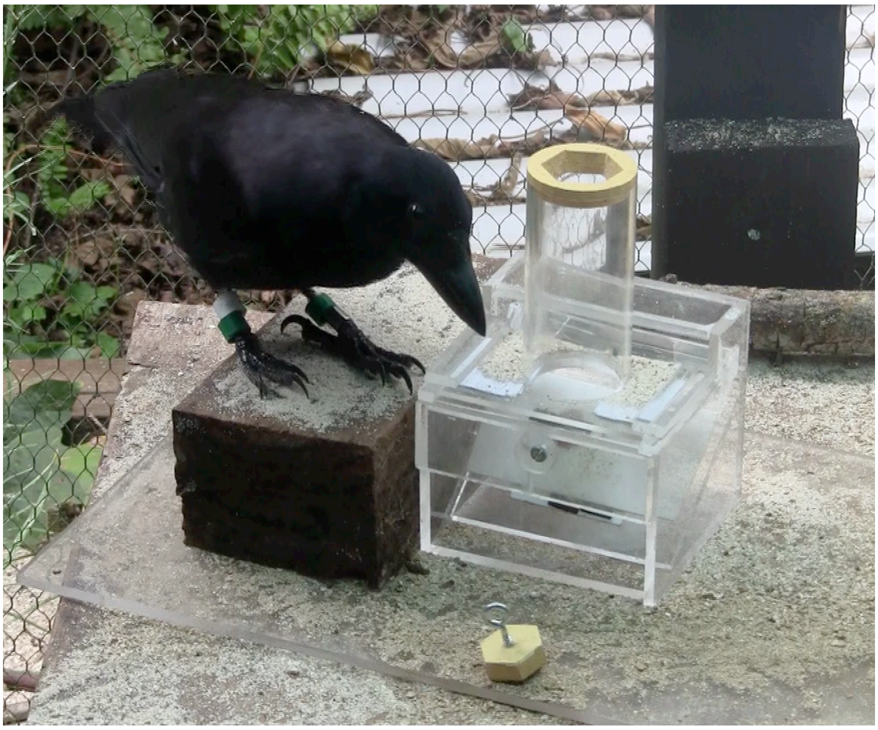

Figure S1: A modified version of the apparatus of Bird & Emery (2009). Only an object of the right shape can be inserted through the cap on the tube to release the reward. Combinations of such boxes with different caps can be used to present multiple new instances of a planning problem. Photo courtesy of Alex Taylor.

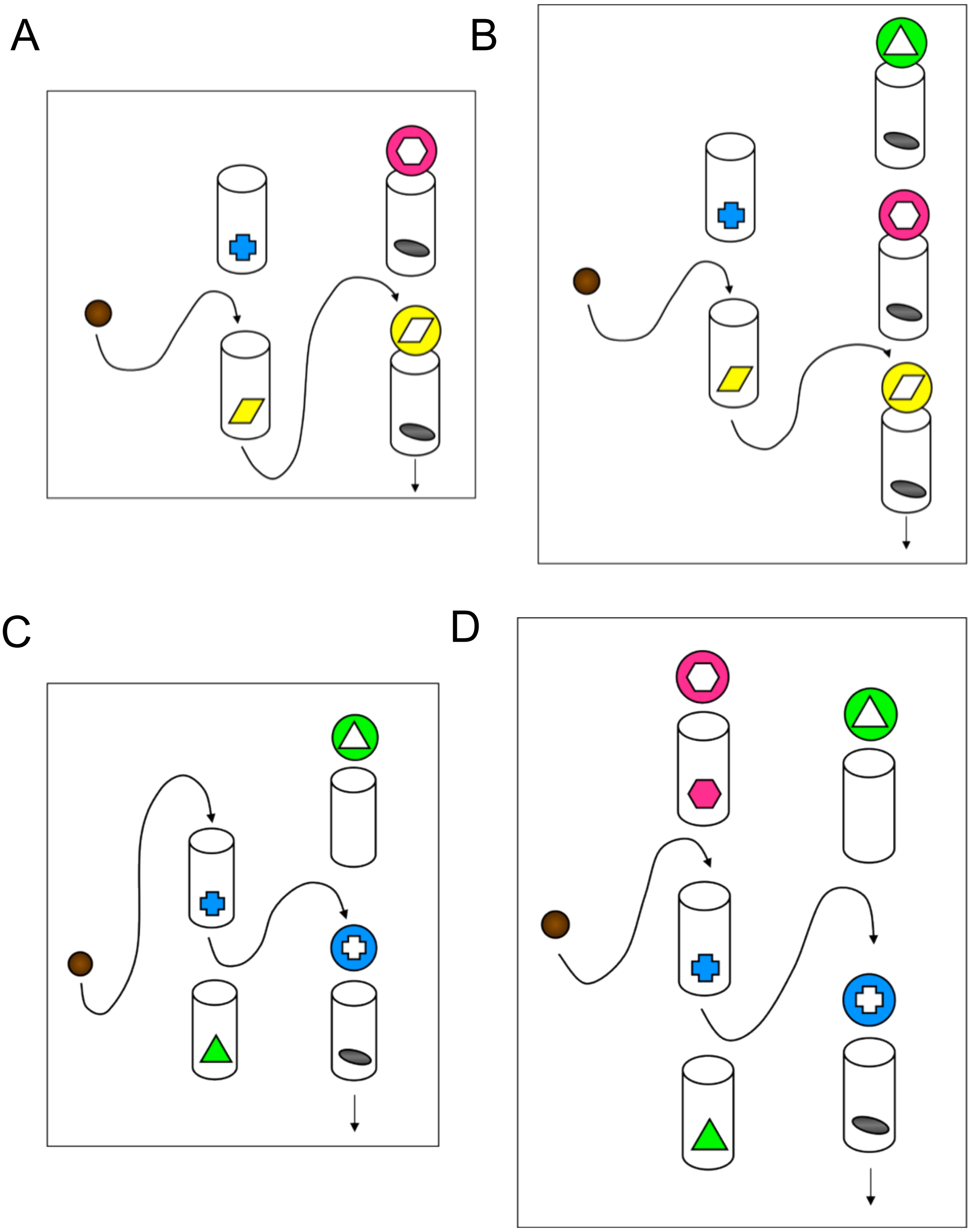

Figure S2: An experiment using multiple variants of the apparatus of Fig. S1 for discriminating between a forward chaining approach to planning and a backward chaining approach. If the subject uses backward chaining, then B (three initial possibilities) should be a greater challenge than A (two possibilities), with little difference between C and D (two possibilities each). Conversely, if the subject uses forward chaining then D (three possibilities) should be a greater challenge than C (two possibilities), with little difference between A and B (two possibilities each).

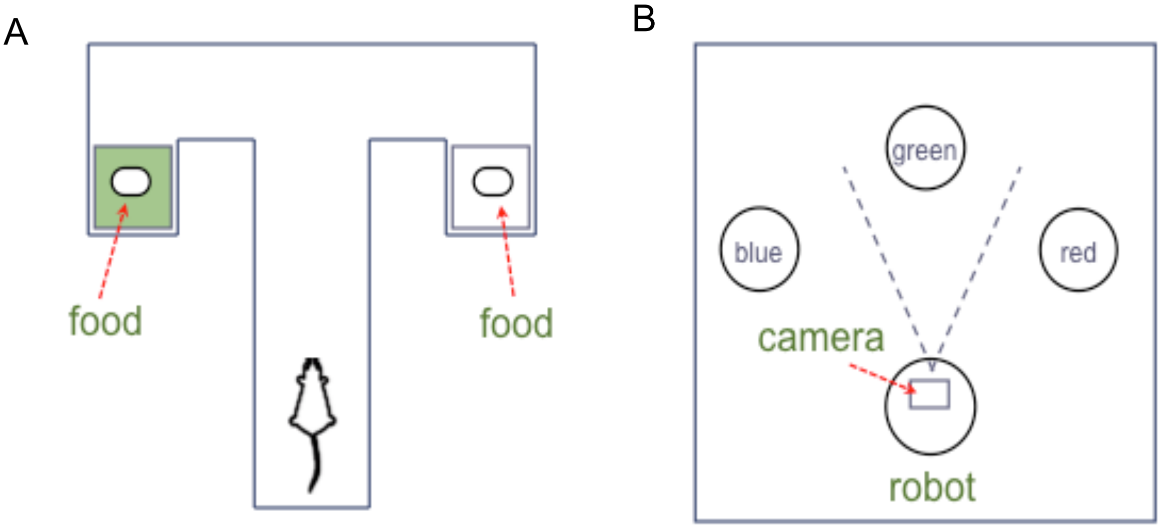

Figure S3: (A) The T-maze experiment of Tolman and Gleitman (1949). The rat explored the maze freely, and was rewarded in both arms (dark- and light-coloured). It was then removed from the maze, and subjected to a shock in a dark-coloured room. Subsequently it showed a preference for the right (light-coloured) arm of the maze, although neither arm was visible when it made its choice, suggesting a capacity to anticipate the effects of actions. (B) A robot analogue of Tolman & Gleitman's experiment (Shanahan, 2006). The robot has an initial preference for turning right when facing green. But it is trained to associate turning right with red, for which it has an aversion. Thanks to an inner rehearsal mechanism, it can anticipate the effect of turning right, causing it to overcome its prior preference and turn left instead.