Cogn Comput (2009) 1:29–41
DOI 10.1007/s12559-009-9010-2

# Reverse Engineering the Vertebrate Brain: Methodological Principles for a Biologically Grounded Programme of Cognitive Modelling

**Kevin N. Gurney**

**Abstract** How are we to go about understanding the computations that underpin cognition? Here we set out a methodological framework that helps understand different approaches to solving this problem. We argue that a very powerful stratagem is to attempt to 'reverse engineer' the brain and that computational neuroscience plays a pivotal role in this programme. En passant, we also tackle the oft-asked and prior question of why we should build computational models of any kind. Our framework uses four levels of conceptual analysis: computation, algorithm, mechanism and biological substrate. As such it enables us to understand how (algorithmic) AI and connectionism may be recruited to help propel the reverse-engineering programme forward. The framework also incorporates the notion of different levels of structural description of the brain, and analysis of this issue gives rise to a novel proposal for capturing computations at multiple levels of description in a single model.

**Keywords** Methodology · Levels of analysis · Computation · Cognition · Basal ganglia · Brain architecture · Computational neuroscience

## Introduction

There are several ways we might go about understanding a cognitive agent. One approach is simply to observe its behaviour—how it responds to various inputs and interacts with its environment—and to try and construct another, artificial agent which displays similar behaviour. An alternative to this constructive technique starts by *deconstructing* the agent at the physical level. That is, we 'look inside' to discover the agent's physical makeup and establish the interconnection between, and function of, its component parts. If the agent is a biological one, deconstruction will involve gathering data on the structure of the brain; this is the remit of neuroscience. Subsequent *re*construction of the agent's cognitive abilities, based on the neuroscientific data obtained in deconstruction, will result in a deep understanding of these abilities, and involve the building of quantitative models; this is the remit of computational neuroscience. The combination of deconstruction and functional reconstruction constitute the process of 'reverse engineering' the brain.

Reverse-engineering biological systems offers a powerful paradigm for understanding cognition. Nature presents us with the opportunity of finding solutions to a plethora of computational problems that define cognition and which, crucially, work in synergy with each other—an issue which is revisited in the section "Algorithms and the brain". While we promote computational neuroscience as a key activity in this programme, this does not imply that other, less biologically grounded computational approaches (including AI and connectionism) are redundant. Rather, these activities have key roles to play in developing the reverse-engineering programme, and these roles will be better appreciated when we have developed the principled methodology for doing computational neuroscience. The paper therefore deals largely with an exposition of this methodological framework and exploring its implications.

However, before this, we take a step back and try to answer a prior question: Why should we build computational models of any kind? Presumably, most readers of this journal will accept the rationale for a programme of

K. N. Gurney (✉)
Department of Psychology, Adaptive Behaviour Research Group, University of Sheffield, Sheffield S10 2TP, UK
e-mail: k.gurney@shef.ac.uk

this kind without question. However, there are many experimentalists in the biological sciences who question its validity. It is therefore worth trying to address their concerns. Only then can we hope for a rich and productive dialog between experimentalists and modellers in the study of cognition.

## The Need for Computational Modelling

Consider the simple neural network model shown in Fig. 1a. It consists of two coupled leaky-integrator model neurons 1 and 2, with neuron 1 receiving a current pulse as well the output of its neighbour. Such neural models are typical of those used in rate-coded models of brain systems, albeit usually in much more complex networks. The diagram in Fig. 1a is beguilingly simple. Treated as a box-and-arrow system with little or no quantitative characterisation, we might be seduced into thinking we can predict the result of the current injection experiment (Fig. 1c perhaps?). However, a proper computational treatment tells us otherwise.

Thus, suppose each neuron is described by an equation of the form $\tau dV/dt = -V + F$, where $V$ is a 'membrane potential' variable, and $F$ is a forcing term. For neuron 2, $F_2 = w_{12}\ y_1$, where $y$ is the output of neuron 1, and for neuron 1, $F_1 = w_{21}\ y_2 + I$ where $I$ is an injected 'current' (shown in the circle). The outputs are given by a piecewise linear function, $h$ (Fig. 1b) where $y = h(V)$. We now ask the question: What is the output of neuron 1 when a rectangular current pulse is input to the network? Four possible results are shown in Fig. 1c–f. They include a simple exponential rise or fall to equilibrium (panel c), oscillatory behaviour (panel d), a rebound phenomenon after the current is turned off (panel e), and 'latching' in which neuron 1 remains active after the pulse has completed (panel f).

The salient point here is that the outcome depends enormously on the network parameters (inter-neuron weights and characterisation of $h(V)$); we cannot predict what the outcome will be by 'doing the simulation in our heads'. The same argument will apply to most 'box-and-arrow' models, or qualitative analyses of neural systems.

Having built a successful computational model, what can it tell us (if anything) about the underlying target system? There is widespread skepticism in much of the neuroscience community about the relevance of modeling. Thus, De Schutter [9] has recently noted that "More than two decades after the declaration of computational neuroscience as a subfield [47] we must conclude that its impact on mainstream neuroscience remains limited and, in particular, most neuroscientists deny theory a strong role in their scientific approaches". Indeed, this problem is not unique to computational neuroscience and appears to be prevalent in much of the non-physical sciences including social science and economics. It was in this arena that Epstein [12] recently provided an extremely well argued case for modelling, but his arguments could apply equally well to computational neuroscience and cognitive modelling in general.

Epstein [12] lays out 16 'reasons to model' *other than prediction*. This is important because it is often thought that the sole purpose of building a quantitative (computational) model is to predict the outcome of future experimental manipulations. While prediction may prove a strong test of a model in the future, at the top of Epstein's list is a very powerful case for modelling—namely *explanation*. In the neurosciences, it is generally agreed that there is an ever growing 'data mountain' which needs interpretation and explanation. If a model can provide a

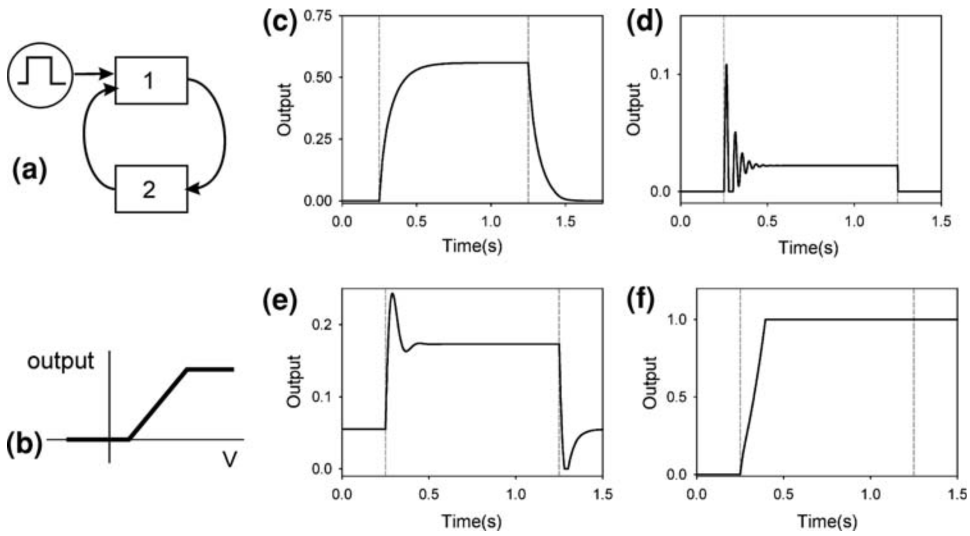

**Fig. 1** Non-intuitive behaviour of even very simple models. **a** A short current pulse is input to a simple network of two leaky integrator neurons. **b** Piecewise nonlinear output function. **c**–**f** Responses of network to a current pulse using different network parameters

mechanistic account for a range of phenomena that were not explicitly used to constrain its construction, then the model has explanatory power and is of real utility. For example, suppose electrophsyiological recordings (possibly from several labs) have been made from neurons in two tightly coupled populations showing behaviours similar to those in Fig. 1. If, as is usually the case, we interpret a rate coded leaky integrator as a model of a population (rather than a single neuron per se), then the model used to derive the results in Fig. 1 would have strong explanatory power.

Finally, it is interesting to note that another of Epstein's reasons to model is to "reveal the apparently simple (complex) to be complex (simple)". Our two-neuron model is a good example of revealing complexity in the apparently simple. We now proceed to describe the methodological framework for computational neuroscience.

## A Principled Methodology to Computational Modelling in the Neurosciences

### Computational Levels of Analysis

How should one go about developing a model of a particular brain system, or of a particular cognitive function? Our answer is based on the proposal by David Marr [34] that brain functions address the solution of computational problems. Further, Marr suggested that these problems decompose (at least in the first instance) into three levels of analysis. At the top level is a description of 'what' is being computed and 'why'—the computational task. This top level is sometimes referred to simply as the 'computation'. In this case, the term 'computation' is used to mean 'function' rather than the act or process of computing. At the next level we describe 'how' the computation is carried out in terms of an algorithm and any associated representations. Finally we specify 'where' the computation is carried out—which brain system implements the computation. This scheme, therefore, enjoins us to specify the cognitive computation as precisely as we can, before proceeding to detail an algorithm and implementation.

Marr's original example [34] provides a very clear illustration, albeit outside the remit of cognitive modelling. Consider the computation of the bill in a supermarket with a cash register. In answer to the top level question of 'what' is being computed, it is the arithmetical operation of addition. As to 'why' this is being done, it is simply that the laws of addition reflect or model the way we should accumulate prices together when shopping; it is incorrect, for example, to multiply them together. Algorithmically, we use the normal procedure (add digits representing the same order of magnitude and 'carry' any overflow if needed). Further, in cash registers, this will be done in the decimal representation rather than binary (normally encountered in machine arithmetic) because rounding errors are incurred when converting between the everyday (decimal) representation of currency and binary. As for the implementation, this occurs using logic gates made out of silicon, silicon-oxide and metal. Notice that choices at different levels are, in principle, independent of each other. For example, we could have chosen to use a binary representation, and alternative implementations might make use of mechanical machines or pencil and paper. The importance of discovering good representations for solving the problem is crucial. Thus, the use of a positional number system, with a number-base and sequentially increasing exponents (like decimal or binary numbers), is the key to the algorithm used here; algorithms for manipulating the number system of the ancient Romans are far more complex.

As a somewhat more realistic application to cognition, consider the problem of directing our visual gaze using ballistic eye-movements or saccades. While we will leave unanalysed several aspects of this problem, our treatment will highlight the need for expansion of the simple trilevel scheme described above. The specific *computational problem* we focus on is: How do we direct gaze to relevant or 'salient' locations in a visually cluttered environment? We will leave unanswered the related problem of finding how to compute what is salient but, in general, this will be determined by a combination of bottom-up feature information (edges, corners and the like) and top-down task information [7].

One *algorithm* for doing this is shown subsequently.

**Algorithm 1** An algorithm for directing visual gaze to salient points in space

```
divide visual space into a set of small regions, ${R_i}$ centred on $\mathbf{x}_i$
for each $R_i$ do
    assign salience $s_i = S(\mathbf{x}_i)$
end for
find location of maximal salience $\mathbf{x}_{max} = \mathrm{argmax}(S(\mathbf{x}_i))$
direct gaze to $\mathbf{x}_{max}$
```

The *representation* used in the algorithm is the set of the spatially localised saliences $S(\mathbf{x}_i)$). The *implementation* of the algorithm must be done somewhere in the brain, but there is, as it stands, no obvious way of implementing the 'neuron-free' algorithm into brain circuitry. What is needed is another level of analysis which has to do with *neural mechanisms*. Thus, we could propose that the calculation of the maximally salient location $\mathbf{x}_{max}$ is performed by a winner-take-all network (or combinations of such networks over several spatial scales, if need be).

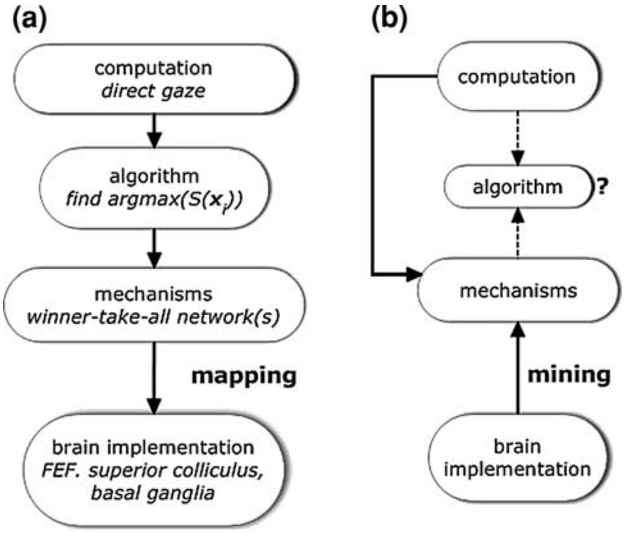

**Fig. 2** Four-level scheme for analysing biological cognitive computation, and methods for using it. **a** Top down: mechanism mapping. **b** Bottom-up: mechanism mining

In general, therefore, we suggest that Marr's tri-level scheme is augmented with an additional, mechanistic level of analysis as shown in Fig. 2a [19].[1]

In our toy example, while the top three levels of analysis have been identified, it remains to show how the abstract neural mechanisms (winner-take-all nets) may be *mapped* onto brain circuits. Figure 2a shows the suggested mapping involves frontal eye fields (FEF), superior colliculus and basal ganglia, all of which have been implicated in gaze control [14, 21, 45]. This process of mechanism mapping—from abstract neural mechanisms onto brain circuits and systems—is the most challenging step in this top-down approach; will the abstract neural mechanism correspond in any simple way with a real biological substrate?

An alternative approach that avoids this problems is illustrated in Fig. 2b. Here, we are still working to understand a well-specified computational problem, but bypass, in the first instance, the algorithmic level altogether. Instead, we *mine* for mechanisms immediately available in the brain circuits supposed to serve the solution of the problem. However, this bottom-up approach is not without its drawbacks because, while we are guaranteed a biologically plausible solution, we are now bereft of an algorithm and so there remains the problem of future *algorithm extraction*; this issue is revisited in the section "Algorithms and the brain".

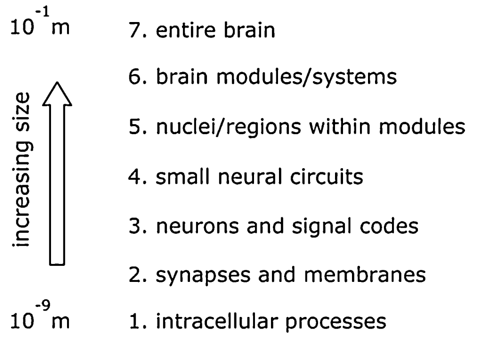

**Fig. 3** Multiple structural levels of description in the brain

Note that much simulation modelling in computational neuroscience uses a bottom-up approach and is indifferent to the existence of algorithms. However, according to our methodological prescription, models should always be cognisant of a top level computation. But what if this is not the case—are such models useful? To answer this, suppose we have a highly detailed and realistic model of an individual neuron or neural microcircuit, say, but are unsure of its overall computation. To the extent that such a model has been extensively validated against its biological counterpart by ensuring similar behaviour, the model is a genuine surrogate for the biological system. As such, it may be subject to a slew of experimental manipulations in an attempt to unravel the mechanisms and, subsequently, the computations, that the circuit or neuron performs. These manipulations may be difficult and time consuming (taking many months) in vivo or in vitro, or (more likely) may be impossible, given the current state of experimental technology. In contrast, in silico, the manipulations of the model and the harvesting of results may be quite rapid (hours or days) given readily available parallel and clustered computing. This leads to the notion that the model is more like an 'experimental preparation', having similar status to the in vitro preparations used by biologists. Such in silico preparations will allow us to perform *high throughput neuroscience* with the goal of discovering the computations the biological substrate performs.[2] At the time of writing, the quintessential example of this approach is the Blue Brain project [33] which is building biologically realistic models of cortical circuits.

## Structural Levels of Analysis

There is another sense in which the brain may be analysed at many levels. This time we refer to the empirical observation that the brain may be described at multiple *structural* levels of description (Fig. 3)

---

[1] In Marr's original formulation of the computational framework, which appeared in an MIT technical report [35], a fourth level was described. However, this was dropped in the more popular account in Marr [34]. Independently, Gurney proposed a four level account in Ref. [15] which was subsequently developed in Ref. [19].

[2] It is often argued that a 'divine gift' of a complete model of the brain would be useless. In the light of the above discussion, however, it would appear this is not true. It may be arduous to unravel the function of all aspects of the model/brain, but this task would certainly be easier than using biological experiments alone.

At level 1 are intracellular signaling processes initiated by neuromodulators (like dopamine). Modelling at this level is the domain of what is now known as *computational systems biology* [30]. Levels 2 and 3 deal with individual neurons. At level 2, patches of neural membrane or single neurons are modelled using, typically, the *Hodgkin Huxley formalism* (see, e.g., Koch [31]). This describes the dynamics of the membrane in terms of the multiplicity of ionic currents it supports. At the next level, we deal only with whole neurons and are more interested in neural firing patterns. Models are often couched in a simplified or *reduced* form—using only two variables—without recourse to a detailed, multi-variable description of membrane biophysics [29]. Also included here are the extremely simplified leaky-integrate-and fire (LIF) neuron models which use only a single variable representing the membrane potential. At level 4, we consider microcircuits within a single brain nucleus. The example *par excellence* here is the cortical microcircuit that extends over six neuronal layers. At level 5, microcircuits are agglomerated into brain nuclei and, beyond that, into entire functional modules (such as cortical sub-systems, hippocampus, basal ganglia etc.). Models at these levels typically use rate coded neurons (leaky integrators) but may also use LIF spiking neurons. In practice, models may draw on features from adjacent levels and there is nothing special about the division into seven levels described here.

We are now enjoined to use two frameworks of analysis—one structural and one computational; how are they to mesh with each other? One possibility is to suppose that computations are defined at a high (or systems) level dealing with modules, say, and that implementation occurs at the neural and molecular levels. Algorithms (and any other intermediate levels of analysis) then sit somewhere in between. This scheme interprets the two frameworks as somehow running in parallel with each other, with each one somehow mirroring the other. However, we argue that the two frameworks deal primarily with different ontological categories and are therefore best thought of as ‘orthogonal’ to each other. Thus, the four-level computational framework has to do mainly with the ideas and concepts of computation, algorithm, abstract mechanism and representation; the implementation is the only reference to the realm of the physical. In contrast, the structural hierarchy is rooted entirely in the physical and delineates objects typical of certain spatial scales.

We, and others [5], therefore believe it makes more sense to allow the computational framework to become manifest at every structural level of description, a scheme which harmonises more naturally with an orthogonal relationship between the two frameworks. Thus, each level of structural description is a potential seat of computational function, and so it is just as valid to think of computation at the intracellular level—as studied in computational systems biology [30]—as it does at the level of a brain nucleus. This is not to say that computations and supporting mechanisms may not reside at different structural levels. An example of computation at the individual neural level supported by lower level (dendritic and synaptic) mechanisms is provided by Mel’s models of neuronal receptive fields in the visual system [36]. Here, models of neurons in visual cortex are able to show preferential response to oriented features in a translationally invariant way. This occurs under a mechanism in which physically proximal groups of synapse (synaptic clusters) have to work cooperatively in order for them to activate the cell. The particular patterning of synaptic clustering then endows the neuron with its highly specific response to visual input. The key point in the current context is that the cooperativity mechanism relies on voltage-dependent membrane conductances (synaptic input only becomes effective if it occurs on dendrites which are already partially active, and the necessary ‘seed’ activity is provided by other neigbouring synaptic inputs).

This linking of function across structural levels also shows how computations at lower levels are forced to become manifest at higher levels; there is no sense in which, by working at a larger spatial scale, the details somehow disappear. For example, it is often assumed that neurons compute their linear weighted sum of inputs. This computation is then carried through in neuron models at the circuit and system level. However, as a result of analysis at the membrane and synaptic level in individual neurons, it is apparent that inhibition often acts more divisively rather than subtractively [31]. This fundamental computation, which is an emergent property of biophysical properties of the neuronal membrane, does not become invisible as we ‘zoom out’ and go to higher levels of structural description; rather it remains a ‘mechanistic leitmotif’ that runs across all levels.

At higher levels we may incorporate divisive inhibition phenomenologically as follows. If $x^-$, $x^+$ are normalised inhibitory and excitatory inputs respectively, to a neuron, then we can write their shunting inhibitory interaction as $x^+(1-x^-)$. This approach has been deployed effectively in models with simple spiking neurons [24, 27]. Similarly the function of neuromodulators like dopamine is a result of complex molecular processes. However, it may be possible to capture their overall influence in an approximate fashion using phenomenological modelling [23, 48].

In summary then, lower level computations and mechanisms will often manifest themselves at higher levels, and their influence may be captured by phenomenological modelling. Alternative approaches to simultaneously capturing computations at multiple levels are given in the section "Computation at multiple structural levels revisited".

It is instructive now to revisit the computational analysis scheme and ask the question: Are four levels of analysis enough? In mechanism mining, the validity of assigning a computational hypothesis to a target brain system may be strengthened by incorporating more biological constraints derived from the target system. This extra detail may be at the same, or lower, structural levels of description than some initial model. Thus, if after this biologically constrained *mechanistic enrichment* the ensuing model still performs the required function, then this is taken as corroborating evidence for the computational hypothesis being applicable to the target neural substrate. The evidence is even stronger if the computational ability can be quantified, and the enriched model performs more successfully. In addition, a more detailed model will usually have more explanatory power by making contact with a wider variety of neuroscientific data. On the other hand, if the more realistic model fails to display the required function, we should reconsider the computational identification for the target brain system.

In general then, there may be a range of possible neuronal models of varying complexity, between the abstract neural mechanistic level and the underlying biological systems. However, all such models are qualitatively similar (neural models of some kind) and their number and complexity is not pre-determined. We therefore choose not to finesse Fig. 2 to explicitly accommodate them, so 'four levels are enough'. These ideas are illustrated further in the 'case study' in the next section.

## A Case Study: Action Selection and the Basal Ganglia

We now illustrate the methodologies outlined above in the light of our own work in modelling the basal ganglia at several levels of structural description. The basal ganglia are the largest group of subcortical structures in the human forebrain and have a critical influence over movement and cognition. The basal ganglia have been implicated in a wide range of processes, including perception and cognition (including working memory), and many aspects of motor function. However, one recurring theme [10, 37] is that they are associated with some kind of selection processing. Our work has developed this idea of selection as a unifying computational theoretical framework for understanding basal ganglia function [44]. Thus, we proposed that the main role of the basal ganglia is to solve the problem of *action selection*—the resolution of conflicts between functional units within the brain that are in competition for behavioural (or cognitive) expression.

In this scheme, functional command units send 'action requests' to the basal ganglia in the form of efferent copies of their encoding of action, and the basal ganglia acts as a central 'selector' or 'switch' mediating the competition for these action requests to be expressed (Fig. 4). Within the basal ganglia, these requests are sent through discrete information streams or *channels* which interact within selective or competitive processing mechanisms. Those requests with the largest overall activity or *salience* 'win' these competitions, resulting in depression of activity in the corresponding basal ganglia output channels. This, in turn, results in action selection as follows. Basal ganglia output is inhibitory and is normally active. The output channels form return loops, via thalamus, with the original functional units that made the action requests. On the winning channels, therefore, there is *disinhibition* of the target thalamo-cortical circuits, allowing them to be active, thereby enabling their behavioural expression.

Having proposed a computational function for the basal ganglia, we then proceeded to perform a bottom-up modelling study at the systems level of structural description. This, therefore, raises the question of what mechanisms can be mined from the anatomy that might support selection. Figure 5a shows some of these and indicates a simplified and partial view of basal ganglia anatomy in cartoon form (for a recent review of basal ganglia anatomy and physiology, see [43]).

The main input nucleus in the basal ganglia is the *striatum*. This contains a complex microcircuit with several varieties of interneurons which may support competitive processing. In addition, the main population of projection neurons show a bimodal (up/down state) behaviour that may serve to filter weak action requests. The output nuclei in primates are the internal segment of the globus pallidus

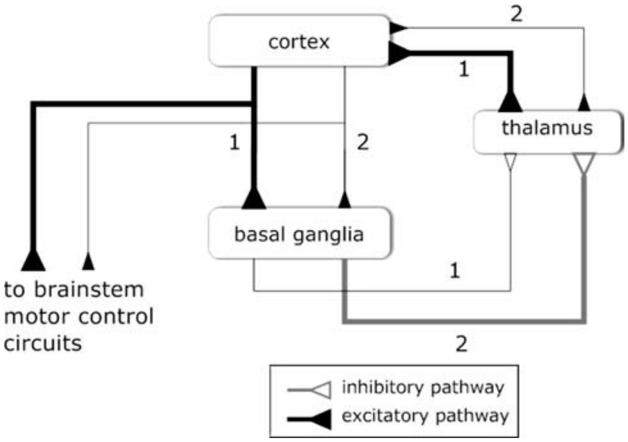

**Fig. 4** Basic action selection mechanism for the basal ganglia. Two action channels (labelled '1' and '2' around the circuit) are shown in competition. *Thick/thin lines* indicate strong/weak signal strengths, respectively

(GPi) and substantia nigra pars reticulata (SNr). These contain lateral inhibitory connections that may support competitive processing. Finally, there is a system wide circuit formed from focussed (‘intra-channel’) inhibition from striatum to the output nuclei, and diffuse (‘cross-channel’) excitation from another input station—the subthalamic nucleus (STN).

At the systems level, we focussed initially on this latter mechanism, which constitutes a feedforward, off-centre on-surround network. It is illustrated in more detail for two channels in Fig. 5b. The polarity of the centre-surround scheme ensures a ‘winner-lose-all’ network, which is just what is required in the circuit of Fig. 4 in order to use release of inhibition as a means of gating actions in target structures.

While the circuit shown in Fig. 5b can, in principle, perform selection, it is not robust against widely varying signal levels (excitation can predominate without careful tuning of weights). However, the correspondence between this circuit and the basal ganglia is also not robust because our description of basal ganglia anatomy has, so far, been somewhat simplified. We now ask: What happens under a mechanistic enrichment at the systems level, when the full basal ganglia circuit is used?

The basal ganglia also contain a nucleus—the external segment of the globus pallidus (GPe)—which receives input from striatum and STN, and which projects only internally to other basal ganglia nuclei. The striatum is also divided into two populations of projection neurons: one which projects primarily to the output nuclei (as shown in Fig. 5), and one projecting preferentially to the GPe. Further, the two populations are distinguished by their preference for dopaminergic receptor types (D1 or D2). We constructed a model of the basal ganglia using the full anatomy as constraint, together with a simple phenomenological model of dopamine effects at D1 and D2 receptors [17] (see Fig. 6).

The resulting model was able to show robust selection and switching between actions consistent with the basal ganglia—action selection hypothesis [18]. In the new functional architecture, we proposed that the role of the GPe is to supply control signals to the selection circuit. Indeed, analysis and simulation confirmed that the closed loop formed by STN and GPe acted like an automatic gain

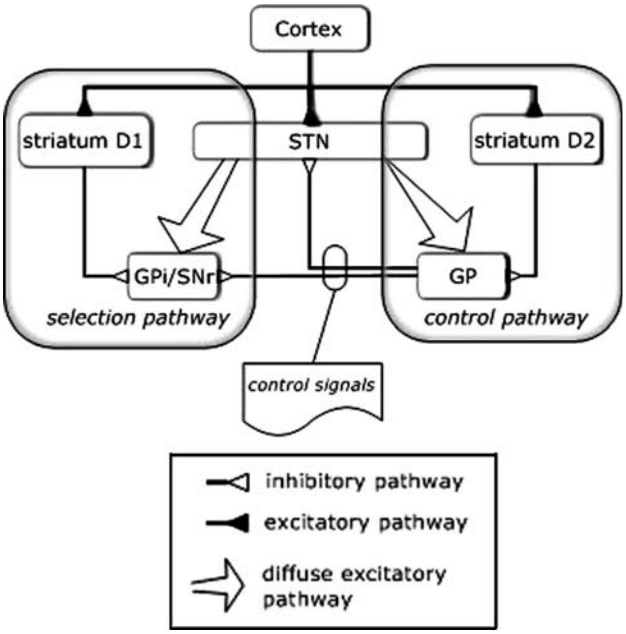

**Fig. 6** Systems level model of the basal ganglia showing new functional architecture with selection and control pathways

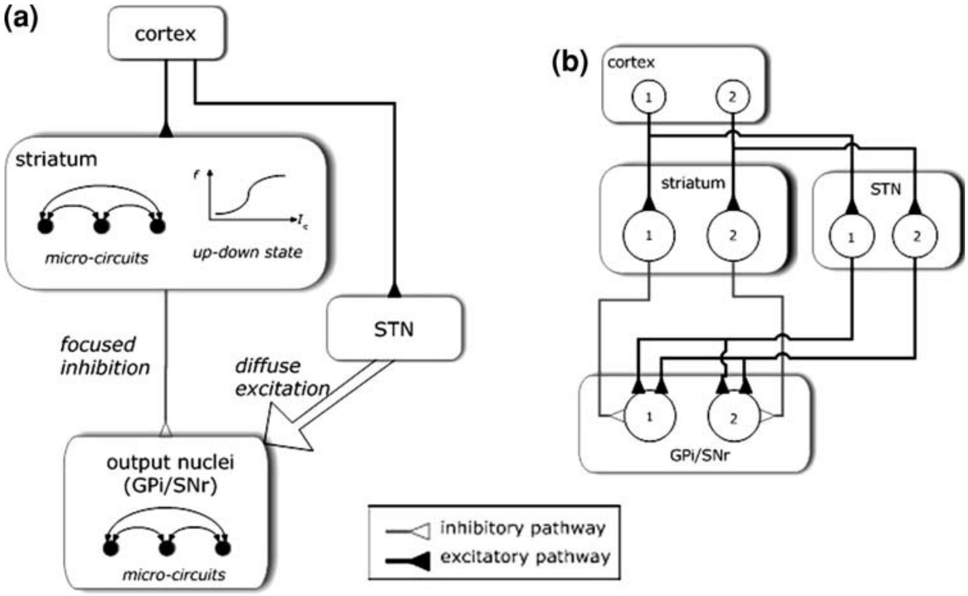

**Fig. 5** Mining for selection mechanisms in the basal ganglia. **a** Mechanisms at systems, circuit and neuronal level. **b** Systems level mechanism constituted by feedforward, off-centre on surround network (two-channel instantiation)

control on STN excitation to ensure the correct operation of the selection circuit over a wide signal range.

Further mechanistic enrichment of the model at the systems level, with additional connectivity intrinsic to the basal ganglia, also increased selection performance [16]. In addition, enrichment at the level of spiking neurons delivered a model which could still carry out a selective function, and account for a wide variety of oscillatory behaviour with the same parameter set [27].

Our work with basal ganglia models has helped shape many of the ideas described in this paper. It will therefore be used again later as an expositional vehicle in subsequent developments and their discussion.

## Computational Neuroscience, Connectionism and AI

We now proceed to consider the broader implications of the framework developed above. In this context, we seek answers to questions like: How do different species of cognitive modelling, such as connectionism and AI, sit within the framework? What is their relationship to computational neuroscience? How can they help in the effort of reverse engineering the brain?

Connectionism deals largely with the study of abstract neural networks. As such it would appear to deal with the mechanistic level of analysis, although principled connectionist models are usually testing a top level computational hypothesis. The existence of a neuronal-free algorithm is not usually addressed, and neither is any mapping onto brain systems (although this is not always the case). This is to be contrasted with computational neuroscience which demands that all four computational levels of analysis be considered. However, the abstraction from direct biological veracity can be a strength if harnessed correctly, because it can lead to discovery of the most general principles operating in the brain that underpin certain cognitive processes.

For example, Hinton and Shallice [22] developed a model of dyslexia with substantial explanatory power. The network was a mixture of feedforward subnets and a recurrent subnet. By making a number of network architectural manipulations (e.g., the use of sparse and fully interconnected feedforward layers, and various placements of the recurrent subnet in the processing hierarchy), they showed that certain aspects of the net's ability to show dyslexic behaviour were contingent only on the existence of a recurrent subnet which can support basins of attraction. The network also illustrates a general feature of many connectionist networks—that multiple psychological phenomena (in this case types of reading error) can be considered to be emergent phenomena resulting from a single system model; multiple patterns of behaviour do not necessarily require multiple paths or subsystems.

In contrast to connectionism, AI deals primarily with computation and algorithm. It has no strong requirement to posit neural mechanisms, much less a biological implementation. The only possible exception here is the study of 'neural networks'. However, we chose to consider this field to be more closely allied with connectionism than symbolic-AI, and it is to this latter specialism we refer to when subsequently using the term 'AI'. However, even if we demand attention be paid only to those algorithms which could be implemented in abstract neural mechanisms, we cannot know a priori which algorithms are suitable candidates for subsequent mechanism mapping. It is therefore worthwhile being as liberal as possible in our algorithm development, and to embrace insights gained directly from studies in AI.

One class of algorithm developed in the AI community which lend themselves to an abstract neural representation are those constructed with reference to directed graphs. The nodes in the graphs may then be interpreted as neural populations and the directed edges as weighted connections between the populations. An example of this class of algorithms is the Bayesian belief networks used to construct cognitive models of the environment. In this scheme, a 'belief propagation' algorithm passes messages between the nodes of a graphical model that captures the causal structure of the environment. Rao [42] shows how these graphs may be implemented in networks of leaky integrator neurons and how these network can then implement sequential and hierarchical Bayesian inference.

Other examples of graph-based algorithms come from the study of visual processing in which a graph or lattice is imposed naturally via the spatial (retinotopic) topography of the visual field. Typically in vision, there is a tradeoff between applying smoothing constraints (visual information tends to be continuous across objects in the world) and segmenting across boundaries (there are different objects and backgrounds). Solutions may be found by constructing Lyaponov or energy function that embody these constraints over a spatial lattice; examples from image velocity encoding and scene segmentation are given in Gurney and Wright [20] and Lee and Yuille [32], respectively.

Even with no obvious graphical interpretation, abstract algorithms find a niche in understanding biological cognition. For example, reinforcement learning (as a branch of machine learning) and, in particular, the temporal difference algorithms have proven to be fertile areas for developing models of reward-driven learning in animals [39, 46].

### Algorithms and the Brain

Notwithstanding the apparent utility of algorithms noted in the previous section, does it make sense to suppose that the brain, in all its intricate complexity, can be described algorithmically? Debate on the relationship between algorithm and brain-and-mind has raged throughout the history of cognitive modelling. It is with some trepidation, therefore, that we now venture a contribution to this debate, but it is an issue which intrudes directly into our four-level computational analysis. Thus, when working in a top-down way, what would be the status of an algorithm which is compelling in its explanatory power, but which cannot be transcribed into an abstract neural mechanism or, even if this is possible, results in a mechanism which maps poorly onto the biology? Conversely, in the mechanism mining approach, are we always guaranteed to extract an algorithm which can capture—at least approximately—the neural mechanism in our original model? We start by looking at some of the historically influential perspectives.

In a seminal article, Putnam [41] presented AI researchers with a stark choice: either it is possible to construct a theory of cognition akin to theories in physics—a single overarching algorithm that will have enormous explanatory power and provide a deep understanding of brain and mind—or AI will be just ’one damned thing after another’—a mixed bag of numerous, and highly diverse mechanisms and software ‘kludges’. The second possibility is a somewhat disparaging description of the proposal by Minsky [38] that mind may be described as a massive collection of semi-autonomous, highly interconnected agents that are themselves mindless. Putnam is skeptical that this approach will prove fruitful but, in contrast, Dennett [8] embraced it dubbing it “Mind as Gadget” and describing it as “an object which one should not expect to be governed by ‘deep’, mathematical laws, but nevertheless a designed object, analyzable in functional terms: ends and means, costs and benefits, elegant solutions on the one hand, and on the other, shortcuts, jury-rigs, and cheap ad hoc fixes.”

It is intriguing to note that modern software architectures have become so complex that their understanding appears to bear striking similarities with the Mind as Gadget. Thus, Booch [3] notes… “Philippe Kruchten has observed that ‘the life of a software architect is a long and rapid succession of suboptimal design decisions taken partly in the dark.’ The journey between vision and ultimate executable system is complex… that path is marked by myriad decisions, some large and some small, some of which advance progress while others represent vestigial dead ends or trigger points for scrap and rework”. Without implying any teleology, this statement could equally apply to the evolution of the brain. Further, Booch coins terms to define software architectures that reflect the mechanism/algorithm debate and our mining/mapping paradigms. “An Accidental architecture… emerges from the multitude of individual design decisions that occur during development, only after which can we name that architecture… An Intentional architecture is explicitly identified and then implemented”. In our language intentional architectures are like the algorithms discovered top down in mechanism mapping, while accidental architectures are like the (potentially very complex) neuronal models built under mechanism mining.

It is not surprising that the brain may be more of an accidental than intentional architecture (Mind as Gadget) because optimising biological (embodied) cognition is a compromise under many conflicting constraints, over millions of years of evolution. A good example of this is provided in visual perception in the primate brain. Thus, there is a massively disproportionate area of visual cortex devoted to the central (foveal) 2–3° of the visual field, with the rest (the periphery) receiving much less neural resource (so-called ‘cortical magnification’ at the fovea). The fovea is, as a result, analysed computationally at very high spatial resolution and along several dimensions (e.g., colour, stereo, object motion, etc.) while information in the periphery is relatively sparsely encoded. If one were designing an optimal vision system *alone*, this may appear to be a rather poor solution. However, the biological resources (brain tissue) that would have to be used in order to represent the entire visual field at the same resolution as the fovea would be enormous; cranial size would be increased by at least an order of magnitude. This is far from optimal in terms of satisfying a multi-objective optimality function which must include other constraints such as the organism’s mobility, and metabolic demand. The biological solution to the problem of perception under embodied constraint such as these is *active vision* (see, e.g., [13]). Here, rapid and accurate eye-movements (saccades) are used to pinpoint a series of foveal targets in quick succession. This enables a representation of the overall scene to be assembled from memory traces of previous views, although the information content therein may be comparatively minimal [4]. The conflicting constraint of cranial size and rapid scene analysis appear to have led to a ‘hacked’ and suboptimal solution for vision, although it is clearly a *sufficiently good* solution for survival.

Pursuing, then, the notion of accidental architecture/Mind as Gadget, what is the status of each of the component gadgets (algorithms)—the so-called “shortcuts, jury-rigs, and cheap ad hoc fixes?” In a seminal paper on connectionism and its relation to symbolic AI, Smolensky [49] supplies one possible answer. Smolensky’s argument is contingent on the designation of a *sub-symbolic* level of analysis which corresponds roughly to our level of abstract neural mechanism. According to Smolensky, the sub-symbolic description of most cognitive processes is the best description we can hope for. We may *attempt* an analysis at a higher, rule-based algorithmic level but any such attempt will be, at best, crude and approximate (certain cognitive processes such as doing mathematics are intrinsically rule based and are exceptions to this rule). The main point is that any high level algorithmic account is an *emergent property* of the sub-symbolic level. As such, it may be *descriptive* of the cognitive ability, but is not a generative account of what *governs* the behaviour.

One answer to the question—Is algorithm extraction guaranteed in mechanism mining?—is, therefore, a clear 'No', for the success of this programme is limited by the extent to which algorithmic regularities emerge from the mechanistic model. Smolensky's account hinges, however, on a particular understanding of sub-symbolic computation; namely that it deals with highly distributed representations in which each sub-symbol (or neural population) takes part in the representation of many high level symbolic objects. This is not always the case, and many computational neuroscience models use localist representations (whose status is less contentious now perhaps, than it was when Smolensky was writing).

This is true, in particular, of the models of basal ganglia described in the section "A case study: action selection and the basal ganglia" in which discrete action channels are encoded by discrete neural populations. This feature of the model makes it a strong candidate for algorithm extraction and, indeed, a successful attempt has been made [2]. The first step was to identify the notion of 'decision making'—normally used by psychologists and neuroscientists in the comparatively narrow context of choice tasks in the laboratory—with that of action selection—normally used by ethologists and roboticists in naturalistic settings. Since action selection is the putative function of basal ganglia, the identification of these two functions implies that the basal ganglia is a decision-making device. Decision making has an extensive theoretical foundation [1] and, in particular, there is an optimal statistical test for decision making with more than two choices the multiple sequential probability ratio test (MSPRT). Bogacz and Gurney [2] showed that many aspects of the anatomy and physiology of the circuit involving the cortex and basal ganglia are exactly those required to implement MSPRT.

Without specifying the process in detail, Fig. 7 gives a flavour of how this was done. Thus, each node in a directed graph describing a subset of the basal ganglia anatomy is associated with an algebraic expression, and each such term is a component in the calculation of the MSPRT (the use of directed graphs here mirrors their utility in mechanism mapping, observed in the section "Computational neuroscience, connectionism and AI"). This style of neuronal computation is quite different from that in Smolemsky's sub-symbolic paradigm. It does not, therefore, suffer from the limitations on the significance of associated algorithms imposed therein. However, further work is required to fully explore the extraction of MSPRT from basal the ganglia. For example, is the algorithm able to incorporate detailed descriptions of basal ganglia microcircuits? It may transpire that, like our systems level case study, such inclusions serve to make MSPRT more robust, or it may be that these circuits represent the need to satisfy other, as yet unseen constraints (compare active vision), so that MSPRT gets 'diluted' or approximated under a series of 'kludges'.

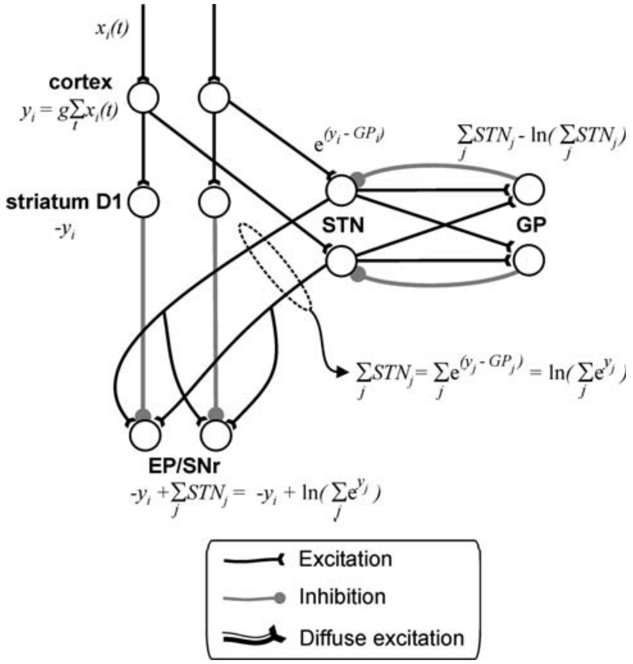

**Fig. 7** Algorithm extraction for the systems level models of basal ganglia; MSPRT decision making

While we have used MSPRT in a computational neuroscience setting, this algorithm has its roots in engineering [11]. It is natural then, to ask, can the interdisciplinary traffic go both ways? Recently, Hussain et al. [28] showed that a control architecture for autonomous vehicle guidance had several features with analogues in the vertebrate brain. This promises a biologically inspired approach to control theory which, in turn, promises discovery of additional brain-relevant algorithms.

Turning now to mechanism mapping—What is the status of an algorithm developed with no reference to the brain (but perhaps with strong explanatory power)? By definition, in this approach, Smolensky's arguments cannot hold sway because we take the position that the algorithm is primary, and that the relevant brain circuits optimise an implementation of the algorithm. However, if we cannot discover sufficiently accurate representation of the algorithm in an abstract neural mechanism, and subsequently, a biological implementation (mapping) of that mechanism, we must ultimately abandon the algorithm as biologically plausible.

In sum, under mechanism mining, algorithms enjoy two kinds of status: on the one hand (after Smolensky), they could be emergent, approximate descriptions of a primary explanatory neuronal mechanism or, having extracted an algorithm, we can choose to think of the mechanism as subsidiary, being merely a route to algorithm implementation in the brain. In contrast, under mechanism mapping, algorithms are, perforce, primary and their validity relies on a successful mapping.

## Computation at Multiple Structural Levels Revisited

In the section "Structural levels of analysis" it was noted that mechanisms at lower levels of description will often manifest themselves at higher levels of description, and that their effects maybe be captured at higher levels by modelling them phenomenologically.

An alternative approach is to build a more detailed model at a level low enough to capture all mechanisms of interest at their native level of description, and which is sufficiently anatomically extensive to cover all large-scale (system-wide) interactions that would be included in a higher level model. This was the approach we took with our large-scale spiking model of the basal ganglia [27]. This model implemented the anatomy in Fig. 6 but used leaky integrate-and-fire neurons calibrated against different neuronal species in the basal ganglia. These model neurons also incorporated additional, mechanistic enrichment dealing with dopaminergic modulation, synaptic input and shunting inhibition (see section "A principled methodology to computational modelling in the neurosciences"). While not as vigorously bottom-up in design as the models in the Blue Brain project [33], the resulting model is sufficiently complex to make it an 'experimental preparation' (see section "Computational levels of analysis"), and it continues to be 'mined' for new insights [26].

While such preparation-like models are valuable, mining them can be challenging because all mechanisms at all levels are universally present on a massive scale. It is as if we are confronted with a high magnification view of a very large piece of brain tissue without the ability to 'zoom out' to lower magnification to see the bigger picture. The approach we want to advocate here is inspired by this metaphor.

Thus, suppose we wish to capture computations at the individual neural level using realistic, conductance-based (compartmental) models. It should be possible to study their effects in a small microcircuit of only a few such neurons, so long as their *signal environment* is similar to that which would occur in a homogeneous model of many thousands of such neurons. The signal environment has two aspects: the raw encoding scheme (point event spikes or firing rates) and the patterning and significance of the signals. In the case of the single neuron models, an approximation to their veridical signal environment may be generated by a network of simplified spiking neurons, so long as they can supply spike trains of the right statistics (mean firing rate, bursting or tonic etc.). This network may be quite extensive containing many thousands of neurons, and can also be studied in its own right. We refer to the small cluster of biologically realistic neurons as a *model core*, embedded into a *model surround* comprising the simplified spiking neuron network.

This *core-surround embedding* scheme may be replicated recursively (in the style of nested Russian-dolls) so that the large-scale spiking neural network (previously a model surround) could become a model core by being embedded into a rate-coded, systems level model surround. The transition from spikes to rates (and vice versa) will, of course, require the construction of novel, hybrid model neurons.

Beyond this rate-spike boundary, small-scale rate-coded models can be easily embedded into larger scale surrounds. It is in this context that we implemented a core-surround model of the Stroop task. In this well-studied psychological task, subjects have either to name the colour of the ink in which words are printed or read the words themselves. The task probes decision and response conflict because the words are themselves colour words like 'RED', 'GREEN', etc., and conflict can occur if the word does not match with its ink colour (e.g., 'RED' in blue ink). In this case, (ink)colour naming causes an increased reaction time.

In our model of the Stroop task [50], the core comprised the biologically constrained model of the basal ganglia [17, 18] described in the section "A case study: action selection and the basal ganglia", together with an extension to include cortico-thalamic processing [25]. This was then embedded into a surround comprising the high-level connectionist model of Stroop processing developed by Cohen et al. [6]. This model implements processing of word and colour information to the point where they form the basis of a decision in the Stroop task. In the original model of Cohen et al., the final outcome of each task trial was determined using a simple, mathematically defined decision mechanism. However, under the action-selection/decision-making identification described in section "Algorithms and the brain", we replaced this by the action-selection mechanism of the basal ganglia, thalamo-cortical looped architecture (Fig. 8).

Modelling, in their entirety, the colour and word processing used in the Stroop task at the same level as the basal ganglia would be a massive undertaking, necessitating careful interpretation of large numbers of cortical areas. Indeed, much of the relevant information may be currently unknown. The connectionist model surround therefore serves as an extremely useful expedient.

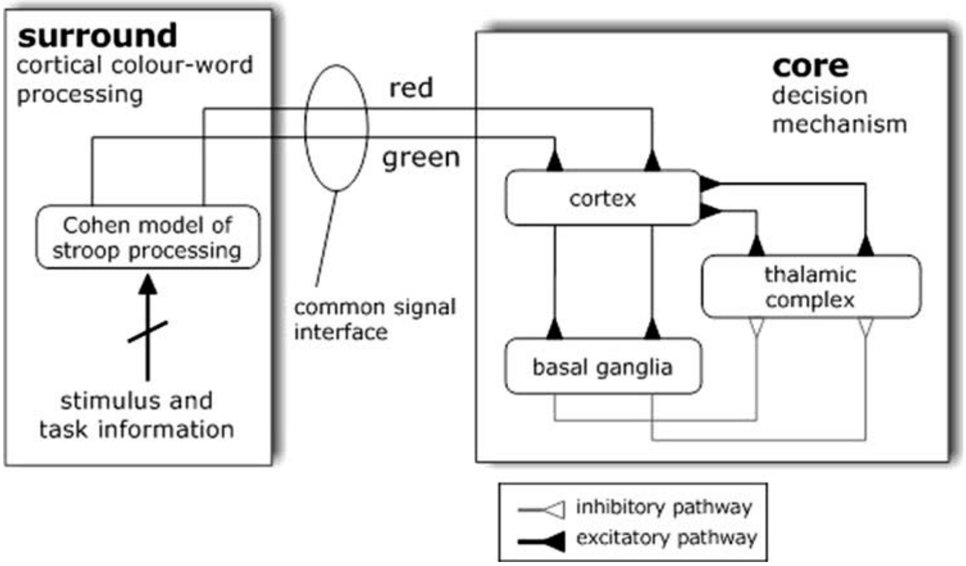

**Fig. 8** Modelling the Stroop task: an example of embedding a biologically constrained core model into a higher level connectionist surround

The key to the success of this project was the observation that the connectionist model supplied output signals which could be interpreted as salience inputs to the basal ganglia. That is, the connectionist model supplied a valid signal environment for the basal ganglia through a common signal interface.

A second application of the embedding scheme is exemplified in our work using models of basal ganglia-thalamo-cortical loops to control autonomous robots [40]. Here, the model surround was the sensory and motor systems that enabled behavioural expression of the robot. These were not biologically mimetic but, nevertheless, supplied a signal environment to the basal ganglia that could test its functionality. Thus, inputs took the form of action requests, and release of inhibition by basal ganglia on motor systems was interpreted as action selection.

In summary, the core-surround embedding (or ‘zoom lens’) approach offers several advantages in working at multiple levels of structural description. First, it promises easier interpretation of the computations being offered at each level of description. Computations at the lowest levels of description may be observed operating veridically in their native mechanistic environment, without recourse to phenomenological approximation. Simultaneously, computations at higher levels will be more easily perceived by being able to ‘zoom out’ to those levels of description. Second, the use of higher level cores obviate the need to model large swathes of the brain at a low level of description when much of the data required to do this properly may simply not be available. Third, the computational resources required for an embedding model will be substantially reduced in comparison with a homogeneous, lower level counterpart.

## Conclusion

We have made a case for quantitative computational modelling as a powerful route to understanding cognition. Within this general strategy we argued that a programme of reverse engineering the brain, by building biologically constrained models using methods in computational neuroscience, holds most promise. Given the enormity of this Grand Challenge, however, we need principled methods of conducting this programme. To this end, we have outlined a four-level framework (computation, algorithm, mechanism and biological substrate) which provides a principled approach to model building. This framework can be used to show the relation between computational neuroscience and other modelling disciplines, such as connectionism and AI, and points the way for a unified attack on the programme by all three approaches. The four-level framework can also encompass working at multiple structural levels of description in the brain (from membranes to systems) and we have proposed a method—the use of core-surround embedding—for working at multiple levels simultaneously.

**Acknowledgements** This work was supported by EPSRC grant EP/C516303/1. I would like to acknowledge all the members of the Adaptive Behaviour Research Group, past and present, who contributed to the work presented here. In particular, I would like to thank Nathan Lepora for reading an early draft of the manuscript.

## References

1. Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. Psychol Rev. 2006;113(4):700–65.

2. Bogacz R, Gurney K. The basal ganglia and cortex implement optimal decision making between alternative actions. Neural Comput. 2007;19(2):442–77.
3. Booch G. The accidental architecture. IEEE Software. 2006; 23:9–11.
4. Chun MM, Nakayama K. On the functional role of implicit visual memory for the adaptive deployment of attention across scenes. Vis Cogn. 2000;7:65–81.
5. Churchland PS, Sejnowski TJ. The computational brain. Cambridge, MA: The MIT Press; 1992.
6. Cohen JD, Dunbar K, McClelland JL. On the control of automatic processes—a parallel distributed-processing account of the stroop effect. Psychol Rev. 1990;97(3):332–61.
7. Connor CE, Egeth HE, Yantis S. Visual attention: bottom-up versus top-down. Curr Biol. 2004;14(19):R850–2.
8. Dennett D. When philosophers encounter artificial intelligence. Daedalus. 1988;117:283–95. Reprinted in ‘Brain Children’ by D.C. Dennett, MIT Press, 1998.
9. De Schutter E. Reviewing multi-disciplinary papers: a challenge in neuroscience? Neuroinformatics. 2008;6(4):253–5.
10. Doya K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? Neural Netw. 1999;12(7-8):961–74.
11. Dragalin VP, Tartakovsky AG, Veeravalli VV. Multihypothesis sequential probability ratio tests—part i: asymptotic optimality. IEEE Trans Inf Theory. 1999;45(7):2448–61.
12. Epstein JM. Why model? J Artif Soc Social Simul. 2008; 11(4):12.
13. Findlay JM, Gilchrist ID. Active vision: the psychology of looking and seeing. Oxford, UK: Oxford University Press; 2003.
14. Girard B, Berthoz A. From brainstem to cortex: computational models of saccade generation circuitry. Prog Neurobiol. 2005;77(4):215–51.
15. Gurney KN. An introduction to neural networks. London, UK: UCL Press (Taylor and Francis group); 1997.
16. Gurney KN, Humphries M, Wood R, Prescott TJ, Redgrave P. Testing computational hypotheses of brain systems function: a case study with the basal ganglia. Network. 2004;15(4):263–90.
17. Gurney KN, Prescott TJ, Redgrave P. A computational model of action selection in the basal ganglia i: a new functional anatomy. Biol Cybern. 2001;84:401–10.
18. Gurney KN, Prescott TJ, Redgrave P. A computational model of action selection in the basal ganglia ii: analysis and simulation of behaviour. Biol Cybern. 2001;84:411–23.
19. Gurney KN, Prescott TJ, Wickens JR, Redgrave P. Computational models of the basal ganglia: from robots to membranes. Trends Neurosci. 2004;27(8):453–9.
20. Gurney KN, Wright MJ. A model for the spatial integration and differentiation of velocity signals. Vision Res. 1996;36(18): 2939–55.
21. Hikosaka O, Nakamura K, Nakahara H. Basal ganglia orient eyes to reward. J Neurophysiol. 2006;95(2):567–84.
22. Hinton GE, Shallice T. Lesioning an attractor network: investigations of acquired dyslexia. Psychol Rev. 1991;98(1):74–95.
23. Humphries MD. High level modeling of dopamine mechanisms in striatal neurons (tech. rep.). Sheffield: Department of Psychology, University of Sheffield; 2003.
24. Humphries MD, Gurney KN. A pulsed neural network model of bursting in the basal ganglia. Neural Netw. 2001;14(6-7):845–63.
25. Humphries MD, Gurney KN. The role of intra-thalamic and thalamocortical circuits in action selection. Network. 2002; 13(1):131–56.
26. Humphries MD, Gurney KN. Deep brain stimulation of the subthalamic nucleus causes paradoxical inhibition of output in a computational model of the “parkinsonian” basal ganglia. Society for Neuroscience Annual Meeting Session 622.9; 2007.
27. Humphries MD, Stewart RD, Gurney KN. A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. J Neurosci. 2006;26(50):12921–42.
28. Hussain A, Gurney K, Abdullah R, Chambers J. Emergent common functional principles in control theory and the vertebrate brain: a case study with autonomous vehicle control. Icann 2008;(2):949–58.
29. Izhikevich. Dynamical systems in neuroscience: the geometry of excitability. Cambridge, MA: MIT Press; 2007.
30. Kitano H. Computational systems biology. Nature. 2002;420 (6912):206–10.
31. Koch C. The biophysics of computation: information processing in single neurons. New York: Oxford University Press; 1999.
32. Lee TS, Yuille AL. Bayesian brain: probabilistic approaches to neural coding. In: Doya K, Ishi S, Pouget A, Rao RPN, editors. Cambridge, MA: MIT Press; 2007. p. 145–88.
33. Markram H. The blue brain project. Nat Rev Neurosci. 2006; 7(2):153–60.
34. Marr D. Vision: a computational investigation into human representation and processing of visual information. New York: WH Freeeman and Co.; 1982.
35. Marr D, Poggio T. From understanding computation to understanding neural circuitry (tech. rep. no. AIM-357). MIT; 1976.
36. Mel BW, Ruderman DL, Archie KA. Translation-invariant orientation tuning in visual “complex” cells could derive from intradendritic computations. J Neurosci. 1998;18(11):4325–34.
37. Mink JW, Thach WT. Basal ganglia intrinsic circuits and their role in behavior. Curr Opin Neurobiol. 1993;3(6):950–7.
38. Minsky M. The society of mind. New York: Simon and Schuster; 1988.
39. Niv Y, Schoenbaum G. Dialogues on prediction errors. Trends Cogn Sci. 2008;12(7):265–72.
40. Prescott AJ, Gonzales FM, Gurney KN, Humphries M, Redgrave P. A robot model of the basal ganglia: behavior and intrinsic processing. Neural Netw 2005;19(1):31–61.
41. Putnam H. Artificial intelligence: much ado about not very much. Daedalus. 1988;117:269–81.
42. Rao RPN. Bayesian brain: probabilistic approaches to neural coding. In: Doya K, Ishi S, Pouget A, Rao RPN, editors. Cambridge, MA: MIT Press; 2007. p. 239–67.
43. Redgrave P. Basal ganglia. Scholarpedia 2007. http://www.scholarpedia.org/article/Basal_ganglia.
44. Redgrave P, Prescott TJ, Gurney KN. The basal ganglia: a vertebrate solution to the selection problem? Neuroscience. 1999;89:1009–23.
45. Schall JD. The neural selection and control of saccades by the frontal eye field. Philos Trans R Soc Lond B Biol Sci. 2002; 357(1424):1073–82.
46. Schultz W, Dayan P, Montague PR. A neural substrate of prediction and reward. Science. 1997;275(5306):1593–9.
47. Sejnowski TJ, Koch C, Churchland PS. Computational neuroscience. Science. 1988;241(4871):1299–306.
48. Servan-Schreiber D, Printz H, Cohen JD. A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. Science. 1990;249(4971):892–5.
49. Smolensky P. On the proper treatment of connectionism. Behav Brain Sci. 1988;11:1–23.
50. Stafford T, Gurney KN. Biologically constrained action selection improves cognitive control in a model of the stroop task. Philos Trans R Soc Lond B Biol Sci. 2007;362(1485):1671–84.