Modeling Natural Action Selection

Proceedings of an International Workshop
Edinburgh, UK

July 2005

edited by
Joanna J. Bryson, Tony J. Prescott, and Anil K. Seth

AISB Press
ISBN 1-902956-40-9

Printed and bound in the United Kingdom.

For additional copies, please contact the Editors:

jjb@cs.bath.ac.uk
t.j.prescott@sheffield.ac.uk
seth@nsi.edu

ISBN 1-902956-40-9

Contents

Invited Presentations 1

Theoretical Perspectives

How a biological decision network can implement a statistically optimal test 2
Rafal Bogacz, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathan D. Cohen

Predicting violations of transitivity when choices involve fixed or variable delays to food 9
Alasdair I. Houston, Mark D. Steer, Peter R. Killeen, and Wayne A. Thompson

Combining action selection models with a five factor theory 16
Mark Witkowski

On compromise strategies for action selection with proscriptive goals 24
Frederick Crabbe

He/she-you-I formalism: A heuristic model of (en)action to make decisions 32
Daniel Mellet-d'Huart

Computational Neuroscience of Action Selection

Forced moves or good tricks in design space? Great moments in the evolution of the neural substrate for action selection 36
Tony J. Prescott

Mechanisms of choice in the primate brain: A quick look at positive feedback 45
Jonathan Chambers, Kevin Gurney, Mark Humphries, and Tony J. Prescott

When and when not to use your subthalamic nucleus: Lessons from a computational model of the basal ganglia 53
Michael J. Frank

Action selection in a macroscopic model of the brainstem reticular formation 61
Mark Humphries, Kevin Gurney, and Tony J. Prescott

Contracting model of the basal ganglia 69
Benoît Girard, Nicolas Tabareau, Jean-Jacques Slotine, and Alain Berthoz

The basal ganglia as the selection mechanism in a cognitive task 77
Tom Stafford and Kevin Gurney

Action selection in subcortical loops through the basal ganglia 84
JC Houk, D Fraser, A Fishbach, SA Roy, LK Simo, C Bastianen, D Fansler-Wald, LE Miller, PJ Reber, and M Botivinick

Cognition, action selection, and inner rehearsal 92
Murray Shanahan

Goal and motor action selection using a hippocampal and prefrontal model 100
Nicolas Cuperlier, Philippe Gaussier, Philippe Laroque, and Mathias Quoy

A computational model of reach decisions in the primate cerebral cortex 107
Paul Cisek

Recognizing invisible actions 113
James Bonaiuto, Edina Rosta, and Michael Arbib

Estimation of eye-pupil size during blink by support vector regression 121
Minoru Nakayama

Agent-Based Modelling

Biorealistic simulation of baboon foraging using agent-based modelling 127
William I. Sellers, Russell Hill, and Brian Logan

Tolerance and sexual attraction in despotic societies: A replication and analysis of Hemelrijk (2002) 135
Hagen Lehmann, JingJing Wang, and Joanna J. Bryson

Collective action selection in social insect colonies 143
James A.R. Marshall

Visual communication and social structure – the group predation of lions 146
Alwyn Barry and Hugo Dalrymple-Smith

Having it both ways – the impact of fear on eating and fleeing in virtual flocking animals 152
Carlos Delgado Mata and Ruth Aylett

Building agents to understand infant attachment behavior 158
Dean Petters

Simulation, emotion and information processing: Computational investigations of the regulative role of pleasure in adaptive behavior 166
Joost Broekens and Fons J. Verbeek

Network-Based Modelling

Routine action: Combining familiarity and goal orientedness 174
Nicolas Ruh, Richard P. Cooper, and Denis Mareschal

Modelling routine sequential action with recurrent neural nets 180
Matthew M. Botvinick

Modelling primate task learning requires bad machine learning 188
Joanna J. Bryson and Jonathan C.S. Leong

Modelling perceptual phenomena using temporal abstraction networks 196
Neil Madden and Brian Logan

Prediction of the behavioural strategy in a chemotaxis search task 203
Manuel A. Sánchez-Montañés and Tim C. Pearce

Symbolic Approaches

Selecting actions and making decisions: Lessons from AI planning 208
Héctor Geffner

Building plans for household tasks from distributed knowledge 215
Chirag Shah and Rakesh Gupta

Innate planning mechanisms 221
Sule Yildrim

Robotics

Ecological integration of affordances and drives for behaviour selection 225
Ignasi Cos-Aguilera, Lola Cañamero, Gillian M. Hayes, and Andrew Gillies

Reinforcement learning of stable trajectory for quasi-passive-dynamic walking 229
Kentauro Hitomi, Tomohiro Shibata, Yutaka Nakamura, and Shin Ishii

Hierarchical reactive planning: Where is its limit? 235
Cyril Brom

Modeling Natural Action Selection

Joanna J. Bryson[1], **Tony J. Prescott**[2], and **Anil K. Seth**[3]
[1]Department of Computer Science, University of Bath, BA2 7AY, UK
[2]Department of Psychology, University of Sheffield, S10 2TP, UK
[3]The Neurosciences Institute, 10640 John Jay Hopkins Drive, San Diego, CA 92121, USA
jjb@cs.bath.ac.uk, t.j.prescott@sheffield.ac.uk, seth@nsi.edu

Action selection is an agent's continuous problem of choosing what to do next. In artificial intelligence, this problem has been addressed with strategies ranging from constructing long chains of intentions that provide provably optimal means of achieving goals to reactive or anytime algorithms that do simple lookups based solely on the external environment. But what does nature do?

These are the proceedings from a multidisciplinary workshop held in Edinburgh in late July of 2005 in association with the IJCAI conference of that year. We dedicated the workshop to advancing the understanding of the behavioral patterns and neural substrates supporting action selection in animals — including humans. Our hope was to create this advance by collecting together examples of good work being done in this area and introducing the authors to each other in a workshop-style context. We also engaged three leading researchers to deliver keynote presentations on different aspects of action selection: Randall O'Reilly (Colorado, USA; Neuroscience), Michael Laver (New York University, New York, USA: Political Science), and Marius Usher (Birkbeck College, London, UK: Psychology). Our aim in bringing together this group of speakers and authors was to tutor each other both on the methods we used to build our models, and on doing and publishing good science using modelling.

We asked for all submitted papers to:

- Reference or describe a model of action selection,
- Reference or describe a data set derived from the actions of living animals or humans, and
- Make direct comparisons between the model and biological data.

Computational models of natural phenomena *are* hypotheses, no different from other hypotheses except that they are particularly well spelled-out and that their implications can be gathered by sampling the output of the model. In other words, hypothesis testing can take place by comparing the behavior of the model to the behavior of the original targets of the model, the animals. We were just as happy to get papers describing data supporting or undermining existing models/hypotheses as we were to get new models of existing data.

All aspects of action selection were acceptable, from single task performance to evolutionary models of behavior, from individual protozoa to human societies.

We received 36 strong papers for our conference. From these, with the help of our reviewers, we chose 14 for presentation (plus the three invited talks), reflecting five technical approaches:

- Theoretical perspectives
- Computational neuroscience of action selection
- Agent-based modelling
- Network-based modelling
- Symbolic approaches and robotics

We also accepted an additional 18 papers in the proceedings, to be presented as posters at the conference. Finally, the remaining places in the workshop were filled with people who submitted short papers and a few people who attended without formally presenting.

We are grateful to SSAISB for publishing these proceedings. At the same time, we hope that this is only the beginning. These are preproceedings — revised versions of the the papers submitted to the workshop (revised in the light of reviews by two to three members of our program committee of the original draft papers), but revisions finished before the meeting took place. We hope at the meeting itself we will all learn a great deal, and that this will be reflected in publications to come.

One note to those not familiar with workshops that publish proceedings. We do not consider these proceedings a substantial archival publication. Some of these papers are summaries of work already appearing in journal articles, and most of the others we expect will be expanded to journal or archival-level conference publications soon, hopefully helped by the participation of the workshop.

This proceedings reflects a snapshot of our field in early 2005, with work from a variety of backgrounds and in many different stages of completeness. We hope that this will help our participants and others prepare similar sorts of work. In addition, we hope these proceedings will draw attention to the excellent work being done in this field, which we aim to to promote through the publication of extended versions of selected papers in a journal special issue.

We are very grateful to IJCAI for their help in organizing this workshop, even though in the end we chose not to be fully affiliated with them. Thanks in particular to Carlos Guestrin,

and to Rob Milne, who will be much missed. Thanks to Natalio Krasnogor of AISB for help with the proceedings, to Myra Wilson and BiroNet / the EPSRC for providing a bursary for student support. Special thanks to Janet Thomas of BiroNet who also provided much useful advice on running workshops. We are very grateful to John Underwood for help with the workshop finances, and to the staff of Edinburgh First for handling many of the local arrangements. Tony Prescott's role in co-organizing the workshop was assisted by grant support from the EPSRC grant no. GR/R95722, and Joanna Bryson was similarly assisted by the EPSRC grant no: GR/S79299/01.

We owe a significant debt of gratitude to Harriet Warburton and the UK's BBSRC, who provided substantial funding for the workshop thus enabling the participation of our guest speakers and one of our organizing committee. She has offered the following statement:

> I am a Programme Manager in the Biotechnology and Biological Sciences Research Council. My areas of interest and responsibility are animal behaviour, neuroscience, cognitive systems (which includes some areas of modelling natural action selection) and animal welfare. Details of BBSRC's full remit and other activities can be found on our website at www.bbsrc.ac.uk.

Finally, we are very grateful to our talented programme committee for their excellent job of reviewing the papers collected in this volume:

Gordon Arbuthnott, University of Edinburgh, UK
Orlando Avila-Garcia, University of Hertfordshire, UK
Christian Balkenius, Lund University, Sweden
Alwyn Barry, University of Bath, UK
Bettina Berendt, Humboldt University, Berlin, Germany
Hagai Bergman, Hebrew Univerity, Israel
Rafal Bogacz, University of Bristol, UK
Olivier Buffet, National ICT, Australia
Lola Canamero, University of Hertfordshire, UK
Angelo Cangelosi, University of Plymouth, UK
Ricardo Chavarriaga, EPFL, Switzerland
Richard Cooper, Birkbeck college, London, UK
Frederick Crabbe, United States Naval Academy, USA
Nathaniel Daw, University College London, UK
Peter Dayan, University College London, UK
Yiannis Demiris, Imperial College London, UK
Peter Dominey, CNRS, France
Kenji Doya, ATR Laboratories, Japan
Jason Fleischer, The Neurosciences Institute, USA
Philippe Gaussier, CNRS, France
Agnes Guillot, University Pierre et Marie Curie, France
Kevin Gurney, University of Sheffield, UK
James Houk, Northwestern University, USA
Karl MacDorman, Osaka University, Japan
Mark Humphries, University of Sheffield, UK
Mark Humphrys, Dublin City University, Ireland
Jeff Krichmar, The Neurosciences Institute, USA
Brian Logan, University of Nottingham, UK
Will Lowe, Trinity College Dublin, Ireland
Jean-Arcady Meyer, CNRS, France
Michael North, Argonne National Laboratory, USA
Peter Redgrave, University of Sheffield, UK
Frank Ritter, Penn State University, USA
Deb Roy, MIT, USA
David Sallach, Argonne National Laboratory, USA
Emmet Spier, University of Sussex, UK
Kris Thorisson, Reykjavik University, Iceland
Myra Wilson, University of Wales, UK

Invited Presentations

1 Michael Laver[1]

Endogenous Political Parties. The spatial model of party competition is one of the dominant paradigms of contemporary political science. Virtually all spatial models of party competition are essentially static: most key model parameters, including the identity of all parties and rules of interaction between them, are set exogenously; the essential solution concept deployed is some form of static equilibrium. However, recent progress has been made with agent-based models that treat party competition as an evolving complex system that may never reach a steady state, (Kollman, Miller and Page 1992; Kollman, Miller and Page 1998; De Marchi 1999; De Marchi 2003; Kollman, Miller and Page 2003; Laver 2005). The central purpose of this paper is to extend the agent-based model of party competition proposed in Laver (2005) to encompass the birth and death of political parties and thereby make the identity of parties in the system an output of, rather than an analyst-specified input to, the model. This is done by modeling party birth as an endogenous change of agent type from citizen to party leader. In order to do this it becomes necessary to model the "memories" of citizens in the system, an issue that has not previously been addressed in agent-based models of party competition, which have hitherto assumed goldfish memories. The birth and death of parties transforms into a dynamic system even an environment where all agents have otherwise non-responsive adaptive behaviors. Substantively, the original purpose of this modeling exercise was to investigate how key system parameters condition the number and identity of political parties in a given system. An unintended but valuable spin-off has been that we are now able to characterize the overall social welfare of the set of citizens, taken as a whole, as a function of party system parameters.

[1]New York University, New York, USA; with Michel Schilperoord, Erasmus University, Rotterdam, Holland.

2 Randall O'Reilly[2]

Toward an Executive without a Homunculus: Computational Models of the Prefrontal Cortex/Basal Ganglia System. The prefrontal cortex has long been thought to subserve both working memory (the holding of information online for processing) and "executive" functions (deciding how to manipulate working memory and perform processing). Although many computational models of working memory have been developed, the mechanistic basis of executive function remains elusive, often amounting to a homunculus. I present an attempt to deconstruct this homunculus through powerful learning mechanisms that allow a computational model of the prefrontal cortex to control both itself and other brain areas in a strategic, task-appropriate manner. These learning mechanisms are based on subcortical structures in the midbrain, basal ganglia and amygdala, which together form an actor/critic architecture. The model's performance compares favorably with standard backpropagation-based temporal learning mechanisms on the challenging 1-2-AX working memory task, and other benchmark working memory and cognitive control tasks. It also makes a number of testable predictions about the contributions of the basal ganglia and prefrontal cortex in various behavioral tasks, several of which have been tested and confirmed.

[2]University of Colorado, Boulder, USA.

3 Marius Usher[3]

Neurocomputational modeling of human decision-making. Decision-making is one of the most common and, at the same time, open-ended and effortful human activities. One source of its difficulty resides with the need to evaluate alternatives whose 'attractiveness' varies on several incommensurable dimensions. Experimental work in human decision-making has also revealed a series of intriguing behavioral patterns that indicate deviations from normative economic theories and which raise a challenge for the development of a theory of human performance. Here I will review some of these patterns, such as loss-aversion and preference reversals under a series of conditions (time constraints, the introduction of contextual information in the choice set, etc). I will then discuss and contrast a number of neurocomputational theories that have recently been proposed to account for these patterns and to explain the cognitive processes that mediate choice-RT and human decision-making.

[3]Birkbeck College, University of London, UK.

How a biological decision network can implement a statistically optimal test

Rafal Bogacz[1], Eric Brown[2], Jeff Moehlis[3], Philip Holmes[2,4], Jonathan D. Cohen[5,6]
[1]Deparment of Computer Science, University of Bristol, Bristol, BS8 1UB, UK
[2]Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA
[3]Department of Mechanical Engineering, University of California, Santa Barbara, CA 93106, USA
[4]Department of Mechanical and Aerospace Engineering, Princeton University, Princeton, NJ 08544, USA
[5]Center for the Study of Brain, Mind and Behavior, Princeton University, Princeton, NJ 08544, USA
[6]Department of Psychology, Princeton University, Princeton, NJ 08544, USA

Abstract

Neurophysiological evidence due to Schall, Newsome and others indicates that decision processes in certain cortical areas (e.g. FEF and LIP) involve the integration of noisy evidence. Within this paradigm, we ask which neuronal architectures and parameter values would allow an animal to make the fastest and most accurate decisions. Since evolutionary pressure promotes such optimality (e.g. in prey capture and predator avoidance), it is plausible that biological decision networks realise or approximate optimal performance. We consider a simple decision model proposed by Usher & McClelland consisting of two populations of neurons integrating evidence in support of two alternatives, and we analyze the dynamics of this model. We show that in order to implement the optimal decision algorithm (sequential probability ratio test) the linearised network must satisfy the following two constraints: (i) it must accumulate the difference between evidence in support of each alternative, as would be implemented by mutual inhibition between the populations; and (ii) the strength of mutual inhibition must be equal to the leak of activity from each population.

1 Introduction

Decision making is a very frequent element of life of humans and animals, and accuracy and speed of the decisions is critical to animal survival. During millions of years of evolution, evolutionary pressure promoted animals whose brains made more efficient decisions. Hence it is plausible that decision circuits in the brain possess architectures and parameters allowing them to implement optimal or nearly optimal algorithms. Therefore, in order to identify the architecture and parameters of decision networks in the brain, it may be informative to ask what is the optimal algorithm for decision making, and what biologically plausible network may implement this optimal algorithm.

This optimality approach is not guaranteed to reveal the true decision network in the brain. But it can produce interesting and counterintuitive experimental predictions, which may be used to test the model suggested by the approach. Furthermore, the algorithm optimally solving a decision problem may uncover (or may inspire) practical computational applications. This report shows how the mathematical analysis of decision processes may help in understanding them and make predictions concerning the architecture of neural networks involved in decision making. In particular, it identifies parameters of the decision making model proposed by Usher & McClelland (2001) under which the computations of the neural decision network are equivalent to an optimal statistical test for decision making (sequential probability ratio test).

In three following sections we briefly review neurophysiology of decision, optimal statistical test, and the model of decision network by Usher & McClelland (2001). Then in Section 5 we identify conditions under which this model achieves optimal performance. Finally, in Section 6 we list other directions in which the theory has been extended.

2 Neurophysiology of decision

The neurobiology underlying decision making has been extensively studied in a task in which monkeys are presented with a visual field of small dots most of which are moving randomly, but a certain fraction of which are moving left on some trials and right on others (Britten et al., 1993). Typically, the animal is trained to respond by making a saccade in the direction of the coherently moving dots. Figure 1a shows schematically the typical patterns of activity observed in area MT of monkeys performing this task (this area is involved in motion processing). When a stimulus with coherent leftward motion is presented, the firing rate of an MT neuron selective for leftward motion is higher than the firing rate of a neuron selective for rightward motion (Britten et al., 1993) (in Figure 1a the grey curve is more often above the black). However, the firing rates for both types of neurons are noisy, hence decisions based on the activity of MT neurons at a given moment in time would

be inaccurate. This reflects the uncertainty inherent in the stimulus and its neural representation.

Figure 1b shows schematically the pattern of activity of neurons in area LIP, which is involved in controlling eye movements. The LIP neurons are believed to integrate the input from MT neurons over the duration of a trial. The decision based on the integrated evidence, namely on activity of LIP neurons after about 0.5s, is much more accurate. This example illustrates that the decision process may be realized in the neural substrate by the integration of noisy information.

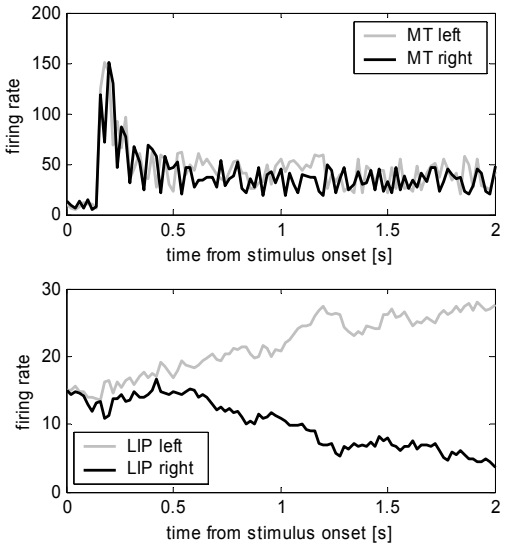


Figure 1. Cartoon of typical peri-stimulus time histograms of neuronal activity during 'the moving dots task'. The figure does not show the actual data, but it is a sketch based on data described by Britten et al. (1993), Shadlen & Newsome (2001), and Schall (2001). Horizontal axes show time from stimulus onset. Vertical axes indicate firing rate. Representative firing rates are shown for stimulus with coherent leftward motion. a) Firing rate of neurons in the area MT: gray line represents a typical neuron selective to leftward motion, and black line for rightward motion (the curves were generated by adding noise to exponentially decaying functions). b) Firing rate of neurons in the area LIP: gray line represents a typical neuron selective for leftward saccades, and black line for rightward saccades (the curves were generated by integrating the difference between curves from panel a and some noise).

3 The decision problem

Let us formalize the decision problem on the basis of the above example. We assume that there are two populations of neurons whose activities provide evidence in support of the two alternative decisions (e.g., corresponding to the two groups of MT neurons in the above example). We denote the activities of the first populations over a given trial by $y_1^1, y_2^1, \ldots, y_n^1$, and of the second population by $y_1^2, y_2^2, \ldots, y_n^2$. Let us assume that the samples y_i^1 come from a normal distribution with mean I_1 and standard deviation c, and samples y_i^2 come from a normal distribution with mean I_2 and standard deviation c. The goal of the decision process is to identify which mean activity, I_1 or I_2, is higher. Note that this is equivalent to asking whether $I_1 - I_2$ is positive or negative, i.e. whether the differences between input samples have positive or negative mean. Let us denote the differences between activities of input populations by $x_i = y_i^1 - y_i^2$.

Within this framework, the question of optimal decision making is the following: For given signal and noise levels I_1, I_2 and c in the input populations, what is the optimal strategy for integration of evidence (e.g., by LIP neurons) that would allow the most accurate and fastest decisions on average, over the course of many trials? More precisely, there are two sub-questions: (i) Which strategy yields the lowest error rate (ER), when a given (fixed) time for decision is allotted, and (ii) which strategy yields the fastest reaction times (RT) for a given ER?

The two questions above refer to optimality in two different conditions under which decision tasks can be run. The first question relates to a decision process in which participants are presented with the stimulus for a fixed duration, at the end of which they are expected to answer, usually on presentation of response prompt, thus constraining their RTs. The second question refers to a decision process in which participants are asked to respond freely, when they are ready, usually being instructed to be as accurate and as fast as possible. We refer to this situation as the *free-response protocol*. We focus only on this protocol.

The answer to the second question, regarding optimality in the free-response paradigm, is provided by the sequential probability ratio test (SPRT) of Barnard (1946) and Wald (1947). In contrast to classical decision procedures in which a previously fixed number of samples is collected before the decision is rendered, SPRT may be applied to continuously accumulating data. A decision is made as soon as a threshold, which depends upon the required accuracy, is reached. Specifically, let H_1 and H_0 denote the two alternative hypotheses, as above, and assume that samples x_i in support of these are drawn at random from two probability distributions with densities $p_1(x)$, $p_0(x)$. In particular, in the case of the decision problem defined at the beginning of this Section, H_1: $p_1(x)$ is a normal distribution with positive mean μ and standard deviation σ, H_0 : $p_0(x)$ is a normal distribution with negative mean $-\mu$ and standard deviation σ. After each sample the ratio of probabilities $p_1(x_i)/p_0(x_i)$ is calculated and the product of these likelihood ratios is accumulated. Observations continue as long as the likelihood ratio lies within the boundaries $Z_0 < Z_1$:

$$Z_0 < \frac{p_1(x_1)p_1(x_2)\ldots p_1(x_n)}{p_0(x_1)p_0(x_2)\ldots p_0(x_n)} < Z_1 \qquad (1)$$

Thus, after each measurement one recomputes the current likelihood ratio, thereby assessing the net weight of evidence in favor of H_1 over H_0. When the ratio first exceeds Z_1 or falls below Z_0, sampling ends and either H_1 or H_0 is respectively accepted; otherwise sampling continues.

SPRT is the optimal test for decision-making on the basis of accumulating noisy data in the following sense: Among all fixed or variable sample decision methods that guarantee fixed error probabilities, SPRT requires on average the smallest number of samples to render a decision (Wald & Wolfowitz, 1948). In other words, for given ER, SPRT delivers the fastest RT.

The SPRT is equivalent to a random walk with thresholds corresponding to alternative decisions. To see this, take the logarithm of both sides of Equation 1:

$$\log Z_0 < \log \frac{p_1(x_1)}{p_0(x_1)} + ... + \log \frac{p_1(x_n)}{p_0(x_n)} < \log Z_1 \quad (2)$$

If we denote the logarithm of the likelihood ratio defined in Equation 1 by I^n, then Equation 1 implies that we iteratively accumulate I^n after each observation:

$$I^n = I^{n-1} + \log \frac{p_1(x_n)}{p_0(x_n)} \quad (3)$$

Let us evaluate the probability ratio for the hypotheses defined earlier in the section. Equation 3 becomes:

$$I^n = I^{n-1} + \log p_1(x_n) - \log p_0(x_n) =$$

$$= I^{n-1} + \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n-\mu)^2}{2\sigma^2}} - \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_n+\mu)^2}{2\sigma^2}} =$$

$$= I^{n-1} + \frac{-x_n^2 + 2\mu x_n - \mu^2 + x_n^2 + 2\mu x_n + \mu^2}{2\sigma^2} =$$

$$= I^{n-1} + \frac{2\mu}{\sigma^2} x_n$$

Thus from the above equation, the SPRT is equivalent to a random walk starting at $I^0 = 0$, and continuing until the logarithm of the likelihood ratio I^n reaches one of the thresholds: $\log Z_0$ or $\log Z_1$. During this random walk the samples (i.e. the differences between the two inputs) are accumulated (a more rigorous and complete analysis of the relationship between SPRT and random walks is given by Gold & Shadlen, 2001).

4 Model of decision network

Figure 2a shows the architecture of an abstract neural network (or connectionist) model for the two alternative decision tasks (Usher & McClelland, 2001). The model includes four units representing the mean activities of neuronal populations: two input units representing populations providing evidence in support of the two alternative decisions (e.g., corresponding to groups of movement sensitive MT neurons from the example in Section 2); and two decision units representing populations integrating the evidence

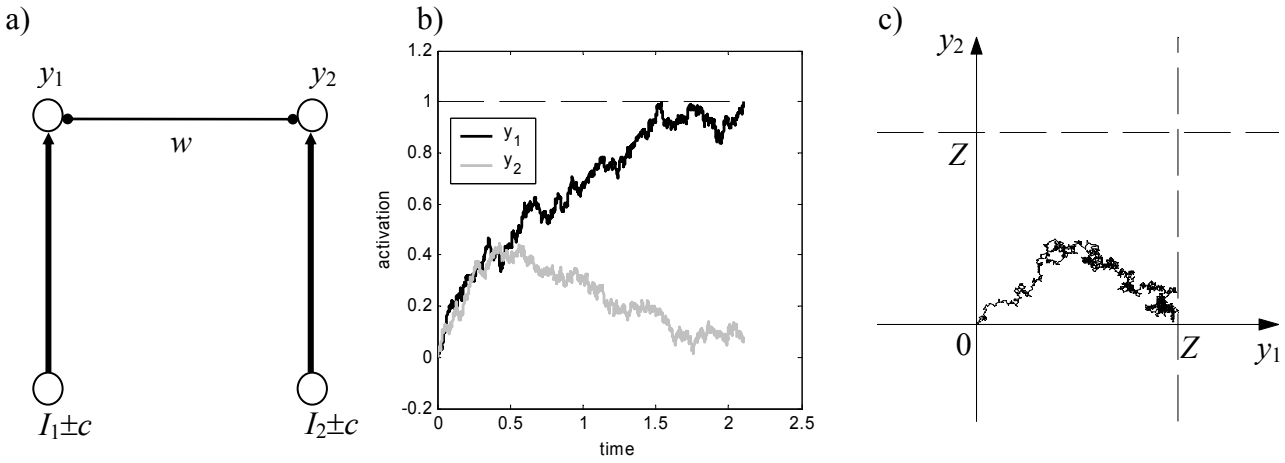


Figure 2. Usher & McClelland (2001) model. a) Architecture of the model: Arrows denote excitatory connections, line with filled circles denotes inhibitory connections. b) An example of the evolution of the model, showing y_1 and y_2 as functions of time. c) The phase- or state space of the mutual inhibition model. Horizontal axis denotes the activation of the first decision unit; vertical axis denotes the activation of the second decision unit. The path shows the decision process from stimulus onset (where $y_1 = y_2 = 0$) to reaching a decision threshold (decision thresholds are shown by dashed lines). The model was simulated for the following parameters: $I_1 = 2$, $I_2 = 1.5$, $c = 0.2$, $w = k = 1.5$, $Z = 1$. The simulations were performed using Euler method with time-step $\Delta t = 0.01$. To simulate the Wiener processes, at every step of integration, each of the variables y_1 and y_2 was increased by a random number from normal distribution with mean 0 and variance $c^2\Delta t$.

(e.g., corresponding to the LIP neurons involved in controlling eye movement).

The decision units are modeled as leaky integrators with activity levels denoted by y_1 and y_2. Each decision unit accumulates evidence from an input unit with mean activity I_j and independent white noise fluctuations η_i of Root Mean Square (RMS) strength c (η_i denote independent Wiener processes). These units also inhibit each other by way of a connection of weight w. Hence, during the decision process, information is accumulated according to:

$$\begin{cases} \dot{y}_1 = -ky_1 - wy_2 + I_1 + c\eta_1 \\ \dot{y}_2 = -ky_2 - wy_1 + I_2 + c\eta_2 \end{cases}, \; y_1(0) = y_2(0) = 0. \quad (4)$$

In the equations above, the term k denotes the decay rate of the units' activity (i.e., the leak) and $-wy_i$ denotes the mutual inhibition. Note that terms $-ky_i$ cause the activity to decay to zero in the absence of inputs to the unit (because if y_i were positive in the absence of inhibition, input, and noise, $\dot{y}_i$ would be negative, and y_i would decrease). The scale of the units' activity is chosen so that zero represents the baseline activity of both units in the absence of all inputs, hence integration starts from $y_1(0) = y_2(0) = 0$. As soon as either unit exceeds a preassigned threshold Z, the model is assumed to make a response.

The state of this model at a given moment in time is described by the values of y_1 and y_2, and may therefore be represented as a point on a *phase plane* whose horizontal and vertical axes correspond to y_1 and y_2; the evolution of activities of the decision units during the decision process may be visualized as a path in this plane. An example is shown in Figure 2c, corresponding to the individual time courses of y_1 and y_2 shown in Figure 2b.

5 Model parameters resulting in optimal performance

As illustrated in Section 4, the behaviour of the model may be visualized by plotting states on the phase plane. Figure 2c shows a representative path in state space: initially the activities of both decision units increase due to stimulus onset, but as the units become more active, mutual inhibition causes the activity of the 'weaker' unit to decrease and the path moves toward the threshold for the more strongly activated unit (i.e., the correct choice).

To better understand the dynamics of the model, Figure 3 shows its *vector fields* for three different parameter ranges. Each arrow shows the average direction in which the state moves from the point indicated by the arrow's tail, and its length corresponds to the speed of movement (i.e., rate of change) in the absence of noise. In Figure 3, as for most other simulations described in this article, we set $I_1 > I_2$, that is, we assume that the first alternative is the correct one (the opposite case is obtained simply by reflecting about the diagonal $y_1 = y_2$).

Note that in all three panels of Figure 3 there is a line (an eigenvector), sloping down and to the right, to which system states are attracted: The arrows point towards this line from both sides. The orientation of this line represents an important dimension: the difference in activity between the two decision units. Note that the evolution *along* the line differs for different values of decay and inhibition, as does the strength of attraction toward the line, and its location in the phase plane. Most of the interesting dynamics determining decisions occur along this line. Therefore, it is easier to understand these in terms of new coordinates rotated clockwise by 45° with respect to the y_1 and y_2 coordinates, so that one of the new axes is parallel to the attracting line. These new coordinates are shown in Figure 3b, denoted by x_1 (parallel to the attracting line) and x_2. The transformation from y to x coordinates is given by (cf. Seung, 2003):

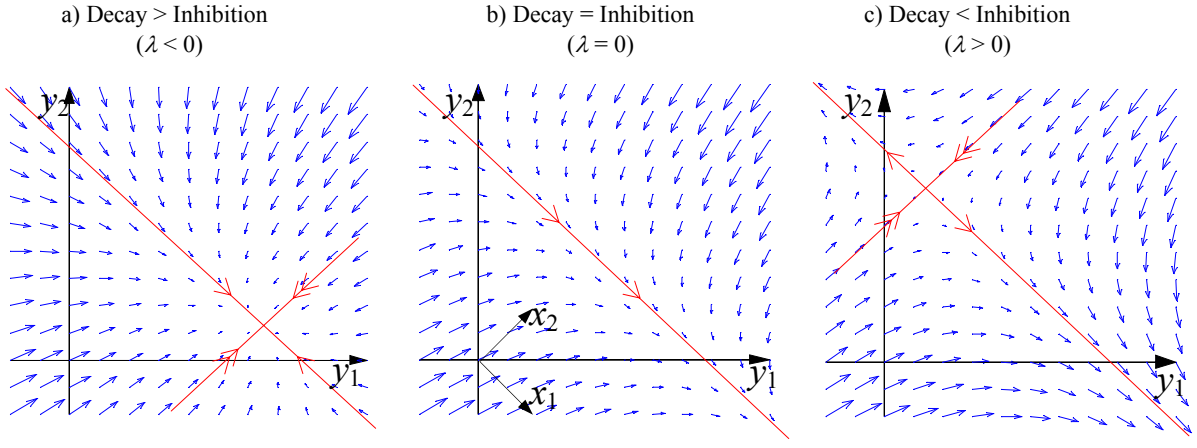


Figure 3. Vector fields for the model. In all plots $I_1 = 2$, $I_2 = 1$. Inhibition (w) and decay (k) have different values in different panels: a) $w = 0.5$, $k = 1.5$; b) $w = 1$, $k = 1$; c) $w = 1.5$, $k = 0.5$.

$$\begin{cases} x_1 = \dfrac{y_1 - y_2}{\sqrt{2}}, \\ x_2 = \dfrac{y_1 + y_2}{\sqrt{2}}. \end{cases} \quad (5)$$

Equations 5 derive from the geometry shown in Figure 3b: x_1 describes the difference between activities of the two decision units, while x_2 describes the sum of their activities. The square root of two in the denominators of Equations 5 is a normalization factor, included to ensure that y and x coordinates have the same scale.

In deciding between two alternatives, it is natural that the *difference* between the activities of the units selective for the alternatives should be a useful descriptor of the decision process (see Section 3). However, the new coordinates do more than merely emphasize this point. They allow us to factor the two Equations 4 that describe the decision process into two decoupled processes, separating the evolution of the difference in the activity of the two units (x_1) from the change in their overall (summed) activity (x_2). If we can show that the latter has minimal impact on the decision process, then we can reduce the description of this process from one that is two-dimensional to a simpler one that is one-dimensional. As we will show, for certain parameters this one-dimensional description reduces to the diffusion model (Ratcliff, 1978).

To transform Equations 4 into the new coordinates, we first calculate the derivative (rate of change) of x_1. Substituting Equations 4 into the first of Equations 5, we obtain:

$$\dot{x}_1 = \frac{\dot{y}_1 - \dot{y}_2}{\sqrt{2}} = \quad (6)$$
$$-k\frac{y_1 - y_2}{\sqrt{2}} + w\frac{y_1 - y_2}{\sqrt{2}} + \frac{I_1 - I_2}{\sqrt{2}} + \frac{1}{\sqrt{2}}\left(c\eta_1 - c\eta_2\right)$$

We assumed earlier that the noise processes for the input units are independent. Since the standard deviation of the sum (or difference) of two independent random variables is equal to the square root of the sum of their variances, the noise process in x_1 may be written:

$$\frac{1}{\sqrt{2}}\left(c\eta_1 - c\eta_2\right) = \frac{\sqrt{c^2 + c^2}}{\sqrt{2}}\eta_{1'} = c\eta_{1'}. \quad (7)$$

In Equation 7, $\eta_{1'}$ again denotes a noise process with mean equal to 0 and an RMS strength of 1. Substituting Equation 7 and the definition of x_1 from Equation 5 into Equation 6, we obtain Equation 8. Following analogous calculations for x_2 , we have:

$$\dot{x}_1 = \left(w - k\right)x_1 + \frac{I_1 - I_2}{\sqrt{2}} + c\eta_{1'}, \quad (8)$$

$$\dot{x}_2 = \left(-k - w\right)x_2 + \frac{I_1 + I_2}{\sqrt{2}} + c\eta_{2'}. \quad (9)$$

Equations 8 and 9 are *uncoupled*; that is, the rate of change of each x_i depends only on x_i itself (this was not the case for the decision units in Equations 4). Hence, the evolution of x_1 and x_2 may be analyzed separately, and in fact each is described by an Ornstein-Uhlenbeck (O-U) process (Busemeyer & Townsend, 1993). In particular, Equation 8, for the x_1 process, involves a drift term proportional to the *difference* between the inputs I_1 and I_2. This process may be stable or unstable, depending upon the relative magnitudes of k and w. Equation 9, for the x_2 process, always gives a stable O-U process (corresponding to attraction to the line in Figure 3), since $-k - w < 0$.

We first consider the dynamics in the x_2 direction, corresponding to the summed activity of the two decision units. As noted above, on all panels of Figure 3 there is a line to which the noise-free state is attracted, implying that x_2 approaches a limiting value as time increases. The rate of this (exponential) approach is $\lambda = k + w$, and it is kept constant in the three cases of Figure 3 by setting $k + w = 2$.

Figure 3 also shows that the dynamics of the system in the direction of coordinate x_1 depends on the relative values of inhibitory weight w and decay k. This dependence is due to the fact that the dynamics of x_1 are described in Equation 8 by a O-U process with coefficient $\lambda = w - k$. When decay is larger than inhibition, then $\lambda < 0$, and there is also an attractor for the x_1 dynamics, as shown in Figure 3a. When inhibition is larger than decay, then $\lambda > 0$, and there is repulsion from the fixed point in the x_1 direction, as shown in Figure 3c. The fixed point is a saddle in this case.

Since $|k+w| > |k-w|$ for *all* parameter values $k>0$ and $w>0$, the average state of the system approaches the attracting line faster (and often considerably faster) than it moves along it (e.g., see Figure 2c). Hence, the decision process divides into two phases: an initial phase in which the activity of both units increases quickly, and there is rapid equilibration to a neighborhood around the attracting line; followed by slower motion along the line, governed by an O-U process in which the difference between the activities of the two units grows as one of them prevails and the other subsides.

Most relevant to the current discussion, when inhibition equals decay the term $(w - k)\, x_1$ in Equation 8 disappears. The vector field for this case is shown in Figure 3b. When inhibition and decay are both fairly strong (as in Figure 3b), the attraction toward the line dominates diffusion along it. Hence, typical paths migrate quickly toward the attracting line and then move slowly along (or near) it.

In this case when inhibition is equal to decay, the position of the system in direction x_1 simply accumulates the difference between the evidence in support of first decision and in support of the second decision, and thus undergoes the diffusion process (Stone, 1960; Lamming, 1968; Ratcliff, 1978). Hence when inhibition is equal to decay, the Usher & McClelland model implements the optimal sequential probability ratio test described in Section 3.

Thus one can expect that the Usher & McClelland (2001) model makes the fastest decisions for fixed error rates when it is closest to the diffusion model, namely when the decay equals inhibition. This is indeed the case, as illustrated in Figure 4.

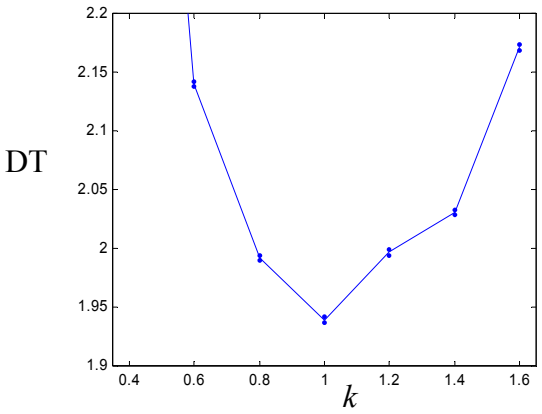


Figure 4. Performance of the model. The following parameters were kept fixed: $I_1 = 1$, $I_2 = 0$, $c = 1$, $w = 1$. Decay (k) is varied (shown on X-axis). Y-axis shows decision time (DT) for the threshold set such that error rate (ER) = 10%. For each set of parameter values, the threshold was increased from zero in steps of 0.01 until the model reached an ER less than or equal to 10%. For each value of the threshold 10,000 trials were simulated. The dots below and above the line indicate the standard error; that is, the standard deviation of DTs across trials divided by the square root of the number of trials (100).

To summarize, when decay is equal to inhibition and both are relatively large, the Usher & McClelland model (2001) approximates the sequential probability ratio test and thus achieves the optimal performance. Thus we predict that in cortical decision network effective decay is also equal to effective inhibition.

6 Further directions

We have extended the theory of neural bases of decision optimization in a number of directions:

- We have shown how more biologically realistic model by Wang (2002) may implement the optimal test (Bogacz et al., 2005).
- We analyzed the values of optimal threshold maximizing the reward rate, which yields a simple relationship between error rates and reaction times, which has been confirmed in a behavioural experiment (Bogacz et al., 2005).
- We analyzed biased decisions in which one of alternatives is more frequent or more rewarded (Bogacz et al., 2005).
- We analysed non-linear version of Usher & McClelland model (2001), and role of gain modulation (Brown et al., 2005)
- We investigated how the optimality generalizes to multiple alternatives (McMillen & Holmes, 2005).

Acknowledgments

This work was supported by the following grants: NIH P50 MH62196, DOE DE-FG02-95ER25238 (P.H.), EPSRC EP/C514416/1, NSF Mathematical Sciences Postdoctoral Research Fellowship (held by J.M.), and the Burroughs-Wellcome Program in Biological Dynamics and Princeton Graduate School (E.B.). We thank anonymous reviewers for very useful comments.

References

Barnard, G.A. (1946). Sequential tests in industrial statistics. Journal of Royal Statistical Society Supplement, 8, 1-26.

Bogacz, R., Brown, E.T., Moehlis, J., Hu, P., Holmes P., & Cohen, J.D. (2004). The physics of optimal decision making: A formal analysis of performaces in two-alternative forced choice tasks. Psychological Review (under review).

Britten, K.H., Shadlen, M.N., Newsome, W.T., & Movshon, J.A. (1993). Responses of neurons in macaque MT to stochastic motion signals. Visual Neuroscience, 10, 1157-1169.

Brown, E., Gao, J., Holmes, P., Bogacz, R., Gilzenrat, M., & Cohen J.D. (2004). Simple networks that optimize decisions. International Journal of Bifurcations and Chaos, in press.

Busemeyer, J.R., & Townsend, J.T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in uncertain environment. Psychological Review, 100, 432-459.

Gold, J.I., & Shadlen, M.N. (2001). Neural computations that underlie decisions about sensory stimuli. Trends in Cognitive Sciences, 5, 10-16.

Laming, D.R.J. (1968). Information theory of choice reaction time. New York: Wiley.

McMillen, T., Holmes, P. The dynamics of choice among multiple alternatives. (under review).

Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 83, 59-108.

Schall, J.D. (2001). Neural basis of deciding, choosing and acting. Nature Reviews Neuroscience, 2, 33-42.

Seung, H.S. (2003). Amplification, attenuation, and integration. Adbib, M.A. (Ed.) The Handbook of Brain Theory and Neural Networks, 2nd edition. Cambridge, MA: MIT Press, pp. 94-97.

Shadlen, M.N., & Newsome, W.T. (2001). Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. Journal of Neurophysiology, 86, 1916-1936.

Stone, M. (1960). Models for choice reaction time. Psychometrika, 25, 251-260.

Usher, M., & McClelland, J.L. (2001). On the time course of perceptual choice: the leaky competing accumulator model. Psychological Review, 108, 550-592.

Wald, A. (1947). Sequential Analysis. New York: Wiley.

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. Annals of Mathematical Statistics, 19, 326-339.

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. Neuron, 36, 1-20.

Predicting violations of transitivity when choices involve fixed or variable delays to food.

Alasdair I. Houston[1], Mark D. Steer[1], Peter R. Killeen[2] & Wayne A. Thompson.

[1]Centre for Behavioural Biology, School of Biological Sciences, University of Bristol, Bristol BS8 1UG, UK

[2]Department of Psychology, Arizona State University, Tempe, Arizona 85287-1104, USA

ABSTRACT

Incentive theory, an established model of behaviour, predicts how animals should choose between alternatives that differ in the amount of food delivered, and the delay until it is delivered. Choice behaviour in such situations may be "irrational", in that it fails to satisfy Strong Stochastic Transitivity (SST). We apply incentive theory within the framework of a standard choice procedure from operant psychology and show that choice behaviour does not satisfy substitutability, and therefore SST does not hold. This occurs because of a change in the context of choice, implicit in the change of experimental conditions necessary to test SST. The predictions from our model are similar to the results of experimental studies of choice behaviour in pigeons. This agreement suggests that behavioural theories may provide insight into other apparent departures from rational behaviour.

1. INTRODUCTION

One of the most commonly cited properties of rational choice is transitivity. In a series of choices between two alternatives, preference is said to be transitive if, from the fact that *a* is preferred to *b* and *b* is preferred to *c*, it follows that *a* is preferred to *c*. Transitivity seems a natural requirement for rationality, in that it produces a ranking of alternatives in terms of preference and eliminates cyclic patterns of preference. A number of cases from the psychological and behavioural ecological domains have been described where decision makers have violated this axiom, thereby behaving in what can be called an irrational manner (Navarick and Fantino 1972; Shafir 1994; Waite 2001b; c.f. Bateson 2002; Schuck-Paim and Kacelnik 2002 where no violations were observed).

When choice is stochastic, rather than deterministic, there are various forms of transitivity (see Fishburn 1973 for a review). Let $p(a,b)$ be the probability of choosing option *a* when faced with options *a* and *b*. Strong stochastic transitivity (SST), with which we are concerned in this paper, requires that

$$\text{if} \quad p(a,b) > \frac{1}{2} \text{ and } p(b,c) > \frac{1}{2},$$

$$\text{then} \quad p(a,c) > \max[p(a,b), p(b,c)]. \qquad (1)$$

SST differs from weak stochastic transitivity in that it puts limits on the magnitude of the preference for *a* over *c*, whereas weak stochastic transitivity simply states that *a* will be preferred to *c*. Violations of both strong and weak transitivity have been shown in grey jays, *Perisoreus canadensis* (Waite 2001b) and honeybees, *Apis mellifera* (Shafir 1994). In this paper we show that a well-established model of choice can account for behaviour that has been interpreted as intransitive.

Tversky and Russo (1969) showed that SST is a necessary and sufficient condition of scalability. Scalability is closely related to the existence of a uni-dimensional model of choice (see Navarick and Fantino 1974; 1975; Houston 1991). Shafir (1994) demonstrated intransitivity of choice in honeybees choosing between artificial flowers varying in nectar volume and length of corolla. Using a conceptually similar experimental procedure, Waite (2001b) observed intransitive choices in gray jays choosing between tubes which differed in the amount of food in the tube and the distance from the entrance to the tube at which the food was placed. Navarick and Fantino (1972; 1974; 1975) found that preferences obtained from a standard choice procedure used in operant psychology (the concurrent-chain procedure) failed to satisfy the requirements of SST, and hence argued that choice probabilities could not be predicted by a uni-dimensional model of behaviour. These results from animals, together with data from humans (e.g. Tversky and Simonson 1993) suggest that the value of an

Corresponding author: a.i.houston@bristol.ac.uk

option depends on the overall context of the choice procedure.

Alterations of the context under which a subject makes a decision have been posited, in different guises, as an explanation for the appearance of irrational behaviours. Schuck-Paim *et al.* (2004) show that unintentional differences in energetic state at the time of decision, resulting from studies' experimental design, can account for some reported violations of regularity (e.g. Waite 2001a; Bateson *et al.* 2002). (Regularity is violated if the proportion of choice for an option is increased after the inclusion of a new alternative in the choice set, Luce and Suppes 1965). This reasoning cannot, however, provide an explanation for the intransitive preferences in which we are interested. Houston (1997) showed that intransitive preferences could arise from a decision process that incorporates a constraint on the accuracy of decision-making. This perceptual error model can predict the qualitative form of the violations of transitivity found by Waite (2001b) and Shafir (1994).

The condition that Houston (1997) showed to be violated was that used by Navarick and Fantino (1972, 1974, 1975):

$$\text{if} \qquad p(a,c) = p(b,c) \tag{2a}$$

$$\text{then} \qquad p(a,b) = \frac{1}{2}. \tag{2b}$$

This condition is part of what Tversky and Russo (1969) call *substitutability*. Following Houston (1991; 1997), we will say that a and b are equivalent in terms of choice against c if they satisfy equation (2a). The other condition for substitutability is

$$\text{if} \qquad p_i(a,c) > p_i(b,c)$$

$$\text{then} \qquad p_i(a,b) > \frac{1}{2}.$$

Tversky and Russo show that substitutability is a necessary and sufficient condition for SST. In going from the premise (2a) to the conclusion (2b) the choice alternative c no longer appears in the context. We note that the absence of c in the choice context may have an effect on the constituent preferences for a and b. In the following section we show that in a standard model of choice it is possible to find many sets of behaviours that satisfy equation (2a) but not equation (2b).

2. THE DELAYED REINFORCEMENT MODEL

We will be concerned with the standard concurrent-chain procedure (see box 1). These schedules were designed to measure preference for environments in which an animal

Box 1.

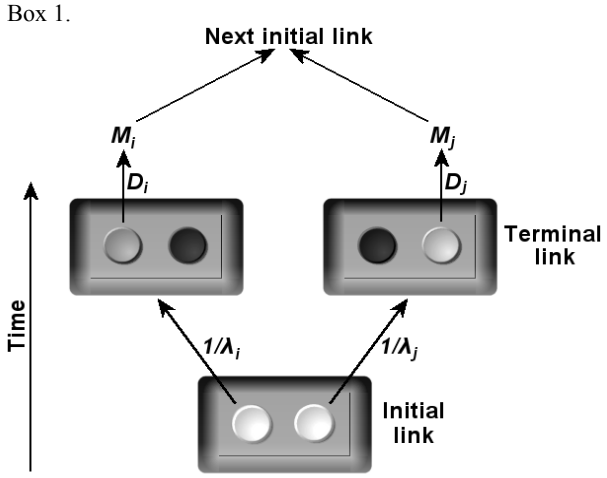


Diagram of the standard concurrent-chains procedure used by Navarick and Fantino (1972; 1974; 1975). During the initial link both keys are white. The keys change colour to signal the terminal link, with one key (the unchosen option) becoming inoperative. Each option is associated with a different terminal link colour. Following a further delay (D) the reward (M) is gained before the initial link recommences.

can work to obtain rewards. To explain this procedure we need to define some of the schedules that are used in operant experiments. On a fixed-interval schedule, an animal receives a reinforcer (e.g. some food) for the first response that it makes on the schedule after a time equal to the schedule interval has elapsed since the last reinforcer. This sort of schedule is usually referred to as a *FI h*, where *h* is the interval. On a variable-interval schedule, the time that must elapse is variable rather that fixed. This sort of schedule is usually referred to as a *VI h*, where *h* is the mean interval.

In a simple concurrent schedule procedure, an animal such as a pigeon is faced with two alternatives, A_1 and A_2. Each alternative has an associated key, which presents stimuli (coloured lights) and measures responses. At the start of the experiment both of these keys are illuminated with a particular colour (e.g. white). The pigeon can peck on either of these keys, which will provide immediate access to food. This procedure may be used to test preferences between different types or amounts of food. The concurrent-chains procedure generalizes this arrangement. Each alternative consists of a VI schedule that, instead of giving the pigeon immediate access to food, gives it access to another schedule that provides food after a delay. The initial VI schedule is referred to as an *initial link* (of the chain) and the schedule to which it gives access is referred to as a *terminal link*. The transition from an initial to a terminal link is marked by a change of key colour. This might be a

change from white to green for A_1 and a change from white to red for A_2. The terminal link determines both a delay until food is obtained and the amount of food that is obtained (for further information see Fantino and Logan 1979). In a given experiment, the schedule on a terminal link may have either a fixed or variable delay. We shall use the subscripts 1 and 2 to identify components in the schedule for A_1 and A_2, respectively. Let $1/\lambda_i$ be the mean delay on initial link i and D_i, be the mean delay on terminal link i (i = 1, 2). The magnitude of reinforcement on the terminal link i is M_i, (i = 1, 2). The relative allocation of responses on the initial link for A_1 is denoted $\alpha(A_1, A_2)$. If N_i is the number of responses made in the initial link of A_i during the course of the experiment then

$$\alpha(A_1, A_2) = \frac{N_1}{N_1 + N_2} \quad . \tag{3}$$

In modelling the above experimental procedure, we use the Delayed Reinforcement Model (DRM) of Killeen and Fantino (1990). The DRM can be derived from Incentive theory (Killeen 1982). We give below a brief description of the DRM, together with a slight extension and modification. Houston (1991) showed that SST can be violated in models of choice between the initial links of concurrent-chains schedules. Houston considered two models: one based on matching, and the other based on the delay-reduction hypothesis. Both models are unable to incorporate variable delays in the terminal links. The DRM includes variable delays and as we show it predicts violations of SST when terminal links involve such delays. It also predicts violations under the conditions considered by Houston (1991).

In the DRM, the choice between two alternatives depends on the product of the rate of access to an alternative and the value of that alternative. The rate of access, r_i, to A_i is

$$r_i = \frac{1}{\frac{1}{\lambda_i} + D_i} . \tag{4}$$

To allow for variable delays in the terminal link of A_i we introduce a set of n_i delays, d_{ij}, j = 1, 2, …, n_i where d_{ij} occurs with probability w_{ij}. The mean delay on terminal link i is

$$D_i = \sum_{j=1}^{n_i} w_{ij} d_{ij} \ . \tag{5}$$

The value of A_i is the sum of the conditioned reinforcing strength, C_i, of the stimuli signalling that terminal link, and the primary reinforcing strength, P_i, of the delayed outcome. The number of responses made on the initial links is

$$N_i = r_i (C_i + P_i) . \tag{6}$$

The conditioned reinforcing strength of a stimulus is proportional to the expected rate of reinforcement that it signals:

$$C_i = p \sum_{j=1}^{n_1} \frac{w_{ij}}{d_{ij}} \tag{7}$$

where p is the constant of proportionality. Other representations of C_i are given in Killeen (1994). The primary reinforcing strength of an incentive is modelled by

$$P_i = q \sum w_{ij} \exp(-d_{ij} /(kTm_i)) \tag{8}$$

where q and k are positive constants and T is the overall time between incentives, given by

$$T = \frac{1 + \lambda_1 D_1 + \lambda_2 D_2}{\lambda_1 + \lambda_2} . \tag{9}$$

For a fixed delay in the terminal links, the analogues of equations of (7) and (8) are

$$C_i = \frac{p}{D_i} \tag{10}$$

and

$$P_i = q \exp(-D_i / kTm_i) \tag{11}$$

respectively. We note that the effect of k is to rescale the magnitude of reinforcement on each terminal link. Let $M_i = km_i$, i = 1, 2 be the rescaled magnitudes of reinforcement. In the following section we present some calculations for the case where the terminal links for each alternative lead to a reinforcement of the same magnitude, but the delay may be fixed or variable. As we are mainly concerned with the effects of different schedules on the terminal links we set $\lambda_1 = \lambda_2 = \lambda$, say. Throughout we set $p = q = 1$. This model gives a good account of the allocation of behaviour to the initial link of concurrent-chains (Killeen 1982); see Grace (1994) for a more general model.

3. ANALYSIS

We now investigate whether substitutability holds when allocation is determined by the DRM. If substitutability does not hold, then SST is violated. The current model

Table 1. The fixed interval schedule *FI*(*h*; *M*, *D*) that results in an allocation of ½ when it is one terminal link and the other terminal link is either (a) a *VI* with mean delay *D* = 23*s* or (b) a *VI* with mean delay *D* = 54*s* when both links result in a reward of magnitude *M*.

(a) *VI* with *D* = 23s, consisting of the following ten intervals, each with probability 0.1:
2.7, 4.1, 5.2, 7.7, 10.1, 16.4, 23.7, 34.8, 52.0, 74.8

M	0.05	0.1	0.5	1.0
FI (*h*; *M*, 23)	8.20	9.43	15.86	17.90

(b) *VI* with *D* = 54s, consisting of the following ten intervals, each with probability 0.1:
2.7, 3.7, 5.7, 9.0, 15.2, 25.3, 43.3, 74.5, 130.0, 228.0

M	0.05	0.1	0.5	1.0
FI (*h*; *M*, 54)	10.67	13.36	28.90	36.92

Table 2. Both terminal links deliver equal reward magnitudes (*M*). For different values of *M,* the *FI* (*h*; *M*, *D*) that results in an allocation of ½ to the initial links is calculated when the FI is one terminal link and the other terminal link is a *VI* with *D* = 23*s*. If SST holds then the allocations to the initial links for *FI*(*h*; *M*, *D*) and *VI* 23*s* will be equal when each is tested against a different option (either *FI 20* or *VI 54*). If the difference between the allocations ($\Delta A_{N(h)}$) ≠ 0, SST is violated.

M	*h*	$\Delta A_{FI\,20}$	$\Delta A_{VI\,54}$
0.10	9.4	0.005	0.031
0.05	8.2	0.015	0.013
0.04	7.2	0.010	0.051

reproduces the findings of Houston (1991) in that substitutability doesn't necessarily hold when fixed-interval terminal links differ in terms of the magnitude of reinforcement and hence the DRM violates SST. Here we look at whether this violation is still obtained when reinforcement magnitude remains constant across the options, but the delay to reinforcement imposed by a terminal link may be fixed or variable. Following Navarick and Fantino (1972) we consider two Variable Interval schedules having means of 23*s* and 54*s*. The intervals that constitute these schedules are given in Table 1; they are the schedules employed by Killeen (1968), the first constituting an approximately rectangular distribution of intervals, and the second an approximately geometric distribution.

Navarick and Fantino (1972) used initial links of 56 or 60*s*. We assume in all cases that the initial links are *VI* schedules with mean interval $1/\lambda_1 = 1/\lambda_2 = 60s$; making alterations to the length of the initial links of just 4*s* makes little difference to the results of our analysis. Each terminal link has the same reward magnitude *M*; we explore the effect of this parameter. When one of the *VI*s given in Table 1 is one of the terminal links, we ask what the value of an *FI* schedule on the other terminal link must be for an animal to have a relative allocation of ½, according to the DRM. Let one terminal link be a variable interval with mean delay *D*. (In fact the mean delay *D* does not completely characterise a schedule, but as we only consider two schedules that have different means, they can be distinguished by their mean delay.) *FI*(*h*; *M*, *D*) is the value of the *FI* on the other terminal link such that the relative allocation is ½ when both terminal link give a reward of magnitude *M*, i.e. α(*VI*(*M*, *D*), *FI*(*h*; *M*, *D*)) = ½. Table 1 gives this *FI* for the *VI* 23*s* and the *VI* 54*s*. In both cases we consider several values of *M*. It can be seen from the table that when one terminal link is the *VI*23, *FI*(*h*; 0.5, 23) = 8.20, which is the value that Killeen (1968) estimated for pigeons.

Navarick and Fantino (1972) tested for substitutability by measuring whether the *VI* and its *FI*(*h*) were equivalent in terms of choice against an *FI*20*s*. We carry out this test with a range of values of the new *FI*. Let the new *FI* have delay *h* and define

$$y(x; M, D) = \alpha(VI(M, D), FI(M, x)) - \alpha(FI(h; M, D), FI(M, x)).$$

If substitutability is to hold, the allocation when one terminal link is one of the *VI*s and the other terminal link is the test *FI* should be the same as the allocation when one terminal link is the *FI* that gives an allocation of ½ against the *VI* and the other terminal link is the test *FI*. In other words, this difference should always be zero. Figures 1a and 1b show that this is not the case for either of the *VI*s, and so out model of preference predicts behaviour that is inconsistent with SST. It is clear that the two allocations will be equal when *x* = *h*, and so *y* will be zero at this point. It can be seen from the figures that y can be zero for other values of *x*.

Navarick and Fantino (1972) produced departures from SST using pigeons, testing for substitutability over a range of reward parameters. Two terminal links were taken to be equivalent if the allocations to both initial links were approximately equal for a pair of options (α(A_1, A_2) = 0.5 ± 0.05); each option was subsequently tested against a third option. Navarick and Fantino arbitrarily counted departures from SST as significant if $\Delta A_{N(h)}$ = |α(A_1, A_3) - α(A_2, A_3)| > 0.05, (*N*(*h*) describes the reward parameters of the third option). Their most striking result was obtained where option 1 = *VI* 23*s* and option 2 = *FI* 7*s*. When these options were each tested against option 3 = *FI* 15*s*, choice was

transitive ($\Delta A_{FI\ 15s} = 0$). In contrast, when option 3 = *VI* 54*s* the resulting choices became strongly intransitive ($\Delta A_{VI\ 54s} = 0.120$). We replicate these experimental findings and show further departures from transitivity over a range of reward values.

Using our version of the DRM if option 1 = *VI* 23*s* and option 2 = *FI*(*h*), if $h < 8.2s$ true equivalence (i.e. $\alpha(A_1, A_2) = 0.5$) is not reached for any value of M. However, at $h = 7.2s$, $M = 0.04$, $\alpha(A_1, A_2) = 0.45$ and by Navarick and Fantino's (1972) criteria this would be counted as equivalence. Under these conditions $\Delta A_{FI\ 15s} = 0.025$ and $\Delta A_{VI\ 54s} = 0.051$, so intransitivity is found in both situations, but it is only apparent to a large enough degree to be counted as intransitive choice using Navarick and Fantino's experimental criteria when option 3 = VI 54s. We have therefore qualitatively reproduced the experimental results using a simple model of choice. Table 2 shows a range of further cases where intransitive choice results from the model, the degree of difference between the initial link allocations, however, is small in each case.

3.1. The causes of stochastic intransitivity

We remarked that in passing from equation (2a) to equation (2b) the context of the choice changes. In general, a possible cause of stochastic intransitivity is that the value of an alternative depends on some aspects of both alternatives (Houston 1991; 1997). In the DRM it can be seen from equation (8) that the primary reinforcement strength of an alternative depends on the overall time between incentives which by equation (9) depends on both D_1 and D_2. To illustrate this, consider choice involving initial links with $1/\lambda = 60s$ and the following three terminal links:

(a) $M = 0.03$, $D = 30s$
(b) $M = 0.036$, $D = 10s$
(c) $M = 0.643$, $D = 90s$

(b) and (c) are equivalent in that when (a) and (b) are the terminal links, the allocation to the initial link leading to (a) is 0.2, and when (a) and (c) are the terminal links, the allocation to the initial link leading to (a) is also 0.2. A necessary condition for SST is that when (b) and (c) are the terminal links, the allocation should be 0.5. In fact, the allocation to (b) is 0.6. The reason for this result is that the overall time T to reinforcement, and hence by equation (11) the *P* value of a terminal link, depends on both terminal links. Thus when (a) and (b) are the terminal links $T = 50s$, whereas when (b) and (c) are the terminal links, $T = 80s$ and when (a) and (c) are the terminal links, $T = 90s$. This example shows that incentive theory does not assign a fixed value to a terminal link; the value of a link depends on the context in which it occurs. It is clear from equation (9) that the effect depends on the fact that the delays on both links influence *T,* which in turn influences *P*. The DRH would not predict an effect in a procedure in which *T* was fixed. Even in the absence of initial links, both D_1 and D_2 influence *T* and hence *P*. Thus violations of SST are theoretically possible even in discrete trial procedures such as that used by Mazur and Coe (1987). As an illustration, in Figure 1c we have taken λ to be quite large (1 sec). The deviations are often greater than when λ = 1/60 (Figure 1b).

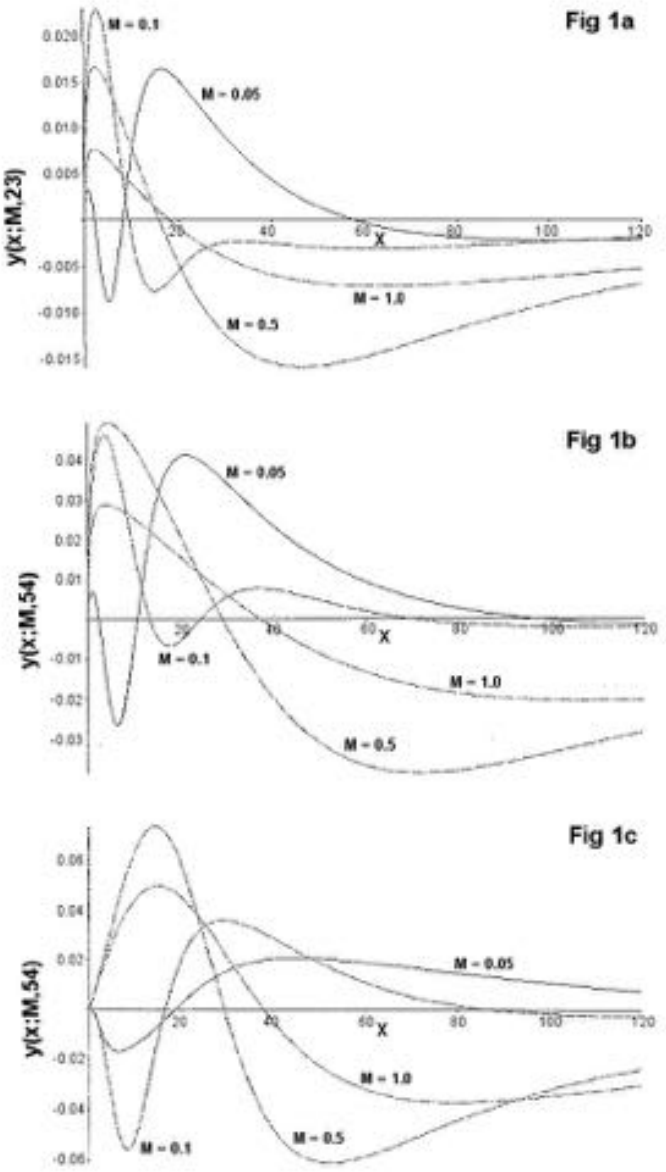


Figure 1. The difference $y(x; M, D) = \alpha(VI(M, D)) - \alpha(FI(h; M, D), FI(x, M))$, where $FI(h; M, D)$ is the *FI* such that relative allocation = ½ when it is one terminal link and a *VI* with mean delay *D* is the other terminal link, both links delivering reward of magnitude *M*. *x* is the duration of the test *FI*. If substitutability is to hold, the $VI(M, D)$ and the $FI(h; M, D)$ must be equivalent in terms of choice against the test *FI*, i.e. $\alpha(VI(M, D)), FI(x, M)) = \alpha(FI(h; M, D), FI(x, M))$ and so *y* should be zero. (a) $y(x; M, 23)$ for $M = 0.05$; $M = 0.1$; $M = 0.5$ and $M = 1.0$. (b) $y(x; M, 54)$ for the same values of *M*. In both cases $1/\lambda = 60s$ and $p = q = 1$. (c) as (b) but with $\lambda = 1$.

4. DISCUSSION

We have shown that the DRM can predict violations of SST on the concurrent-chains procedure, a standard technique for investigating how choice depends on the magnitude of reward and the delay before the reward is obtained. The violations are possible both for *FI*s with unequal reward magnitudes as the terminal links and for *FI*s or *VI*s with equal reward magnitude as the terminal links. As Figure 1 shows, the extent of the violations depends on the parameters in a complex way. Houston (1991) showed for the first case that violations could be predicted from modifications of previous models of choice. The results we present here are the first demonstration of violations in the second case. Whilst the effects predicted by the model are relatively small, we find stronger intransitivity when using similar reward parameters to Navarick and Fantino's (1972) experimental study.

Previously it has been shown that uncontrolled differences in energetic state (Schuck-Paim *et al.* 2004) and errors during foraging (Houston 1997) can produce seemingly irrational behaviour. In this paper we have extended the findings of Houston (1991) to show that intransitivity is also predicted by the DRM. Since the DRM is an established mechanistic account of behaviour, our analysis demonstrates that descriptive models of choice can account for violations of transitivity.

Our analysis of the DRM does not produce departures from transitivity which are as of great a magnitude as found in experimental studies, but the model does not include such factors as differences between individuals, perceptual errors and bias, all of which could interact with a choice mechanism to alter behaviour (e.g. see Grace 1993). It would be interesting and informative for future work to investigate whether incentive theory predicts other seemingly irrational behaviours, such as departures from regularity, when analysed within a framework mirroring other experimental paradigms.

4.1. Context and Choice

The notion that the value of an entity can be dissociated from the context in which it is chosen is one of the many idealizations in science that is correct only to a first order of approximation. The way in which humans frame their choices has profound effects on what they choose (Tversky and Kahneman 1981; Kahneman *et al.* 1982). Decisions made after a loss are different than ones made after a win. Only an Olympian view of value that insists on utility as independent of the state of the user would view such temporal and contextual choices as irrational. There is a large literature on how to assess valuation independently of introspective estimates; finding consilience among the various measures is an unaccomplished task. It may remain unaccomplished, as each of the contexts for assaying preference adds its own character to the choice.

Framing may be conscious or unconscious; in the latter case one speaks more generally of "context effects", sometimes signalled by complaints such as "It seemed like a good idea at the time", "I guess my eyes were bigger than my stomach", or "It looked a lot better on him!". Foraging strategies of less verbal organisms also depend on current resources, contexts and histories. In theory, such context effects can be "internalised" in a model by treating aspects of the choice context as adding value or cost to the object chosen. This is the gambit used in attempting to assay the expected utility of delayed or uncertain goods. Rational models with exponential discounting diminish the value of long-delayed outcomes to negligible values (Ainslie 2001), making prudence irrational and leading behavioural economists to seek other ways of internalising the discount functions.

Self control may be fostered by increasing the salience of the outcomes—listening to preachers elaborate the pleasures of heaven and pains of hell—but our analysis has shown that less dramatic changes in context can have important qualitative effects on preference, in that they undermine one of the standard transitivity axioms of utility theory. The failure of stochastic transitivity seems paradoxical because it is contemplated in a context-invariant manner, as a set of equations and inequalities that are viewed in a moment while reading a paper such as this. The most rational economist will transcend his training to make intransitive choices given the right context. In the experiments we analyse, the context is always changing; the average time in an initial link might rarely be experienced; instead, a sequence of binary choices in a sequence of unique temporal contexts is averaged to represent a dynamic process. The next step for researchers, given the luxury of otherwise-invariant Skinner boxes and organisms at relatively constant levels of deprivation, is to extend the analysis to a real-time model of fluctuations in preferential behaviour as a function of fluctuations in history of exposure (Roe *et al.* 2001). It is only at this level that we are likely to find the invariances that all scientists seek.

References

Ainslie G. (2001) *Breakdown of Will*. Cambridge University Press, Cambridge.

Bateson M. (2002) Context-dependent foraging choices in risk-sensitive starlings. *Anim. Behav.* **64**, 251-260.

Bateson M., Healy S.D. and Hurly T.A. (2002) Irrational choices in hummingbird foraging behaviour. *Anim. Behav.* **63**, 587-596.

Fantino E. and Logan C.A. (1979) *The Experimental Analysis of Behavior*. W. H. Freeman & Co., San Francisco.

Fishburn P.C. (1973) Binary choice probabilities: On the varieties of stochastic transitivity. *J. Math. Psychol.* **10**, 327-352.

Grace R.C. (1993) Violations of transitivity: implications for theory of a contextual choice. *Journal of the Experimental Analysis of Behavior* **60**, 185-201.

Houston A.I. (1991) Violations of stochastic transitivity on concurrent chains: Implications for theories and choice. *Journal of the Experimental Analysis of Behavior* **55**, 323-335.

Houston A.I. (1997) Natural selection and context-dependent values. *Proc. R. Soc. Lond. Ser. B-Biol. Sci.* **264**, 1539-1541.

Kahneman D., Slovic P. and Tversky A. (1982) *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

Killeen P.R. (1968) On the measurement of reinforcement frequency in the study of preference. *Journal of the Experimental Analysis of Behavior* **11**, 263-269.

Killeen P.R. (1982) Incentive theory II: Models for choice. *Journal of the Experimental Analysis of Behavior* **38**, 217-232.

Killeen P.R. (1994) Mathematical principles of reinforcement. *Behav. Brain Sci.* **17**, 105-135.

Killeen P.R. and Fantino E. (1990) Unification of models for choice between delayed reinforcers. *Journal of the Experimental Analysis of Behavior* **53**, 189-200.

Luce R.D. and Suppes P. (1965) Preference, utility, and subjective probability. *In Handbook of Psychology III*, Eds R D Luce, R R Bush and E Galanter. pp 249-410. Wiley, New York.

Mazur J.E. and Coe D.C. (1987) Tests of transitivity in choices between fixed and variable reinforcer delays. *Journal of the Experimental Analysis of Behavior* **47**, 287-297.

Navarick D.J. and Fantino E. (1972) Transitivity as a property of choice. *Journal of the Experimental Analysis of Behavior* **18**, 389-401.

Navarick D.J. and Fantino E. (1974) Stochastic transitivity and unidimensional behavior theories. *Psychol. Rev.* **81**, 426-441.

Navarick D.J. and Fantino E. (1975) Stochastic transitivity and unidimensional control of choice. *Learn. Motiv.* **6**, 179-201.

Roe R.M., Busemeyer J.R. and Townsend J.T. (2001) Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychol. Rev.* **108**, 370-392.

Schuck-Paim C. and Kacelnik A. (2002) Rationality in risk-sensitive foraging choices by starlings. *Anim. Behav.* **64**, 869-879.

Schuck-Paim C., Pompilio L. and Kacelnik A. (2004) State-dependent decisions cause apparent violations of rationality in animal choice. *PLoS. Biol.* **2**, 2305-2315.

Shafir S. (1994) Intransitivity of preferences in honeybees - support for comparative-evaluation of foraging options. *Anim. Behav.* **48**, 55-67.

Tversky A. and Kahneman D. (1981) The framing of decisions and the psychology of choice. *Science* **211**, 453-458.

Tversky A. and Russo J.E. (1969) Substitutability and similarity in binary choices. *J. Math. Psychol.* **6**, 1-12.

Tversky A. and Simonson I. (1993) Context-Dependent Preferences. *Manage. Sci.* **39**, 1179-1189.

Waite T.A. (2001a) Background context and decision making in hoarding gray jays. *Behav. Ecol.* **12**, 318-324.

Waite T.A. (2001b) Intransitive preferences in hoarding gray jays (*Perisoreus canadensis*). *Behav. Ecol. Sociobiol.* **50**, 116-121.

Combining Action Selection Models with a Five Factor Theory

Mark Witkowski
Intelligent Systems and Networks Group
Department of Electrical and Electronic Engineering
Imperial College
Exhibition Road, London, SW7 2BT, United Kingdom
m.witkowski@imperial.ac.uk

Abstract

This paper describes a unifying framework for five highly influential but disparate theories (the five factors) of natural learning and behavioral action selection. These theories are normally considered independently, with their own experimental procedures and results. The framework builds on a structure of connection types, propagation rules and learning rules, which are used in combination to integrate results from each theory into a whole. Exemplar experimental procedures will be used to discuss the areas of genuine difference, and to identify areas where there is overlap and where apparently disparate findings have a common source. The paper focuses on predictive or anticipatory properties inherent in these action selection and learning theories, and uses the Dynamic Expectancy Model and its computer implementation SRS/E as a mechanism to conduct this discussion.

1 Introduction

The overall aim of this paper is to provide a unifying description to encompass and combine five classical and highly influential "theories" of natural action selection and learning. These are the five factor theories. Each held a dominant place in theorizing during the 20th century and was supported by a wealth of meticulously gathered experimental data, but there has been little or no attempt to provide a single framework with which to rationally consider how they might interact.

The problem, in part, arises from the fact that these theories have been treated as largely competitive, at times with considerable animosity being generated between proponents of the differing approaches, or, more often, a tacit isolationism between the different schools of thought.

Such isolationism is surprising, as it clear that individual animals will demonstrate a whole range of behavioral phenomena, each of which might be most satisfactorily described by one or another of the approaches, largely depending on the circumstances the animal finds itself in. It is also very apparent that no single approach explains all animal action selection behavior.

Each factor theory is characterized by the underlying assumption that immediately observable and measurable behavior results from sensations arising from the interaction between the general environment of the organism (including its body) and its sense organs.

The issue under debate was the principles by which that interaction was to be characterized. In itself, expressed behavior gives little indication of which, indeed, if any, of these theories best describes the internal action selection mechanism that gives rise to the observable behavior.

The task, then, is to provide a minimal description of the principles underlying the mechanisms involved that recognizes natural diversity, yet covers the range of phenomena observed. This paper identifies where these mechanisms clearly differ, and where they are apparently different, but can be explained as manifestations of a single type of mechanism, and how these differences may be resolved into a single, structured framework. Given the range and diversity between individual animals and species, there is a fine balance to be struck between highly specific, quantitative, descriptions, trivially refuted due to this natural variation - and untestable generality. This paper attempts such a balance.

The five factor approach described here substantially extends, details and revises the approach to anticipatory learning and behavioral action selection introduced in [Witkowski, 2003]. The approach will be developed in the light of the *Dynamic Expectancy Model* (DEM) [Witkowski, 1998, 2000, 2003] and its actual (C++) computer implementation SRS/E. The analysis in this paper will be performed mainly at the level of the five factor theories, each of which is itself a digest of many exemplar experimental procedures. The paper will call on specific procedures where necessary, and illustrate issues with reference to the DEM and its implementation.

Section 2 provides a thumbnail sketch of each of the five factor theories. Comprehensive descriptions of the five theories can be found in any textbook of natural learning theory (e.g. [Bower and Hilgard, 1981]). Section 3 considers the interface between animal and its environment, and how issues of behavioral motivation might be addressed. Sections 4, 5 and 6 respectively build the arguments for the structural, behavioral and learning

components of the combined approach. Section 7 reconstructs the factor theories in the light of these component parts, and emphasizes the role of the action selection policy map, which may be either static or dynamic. Section 8 describes an arbitration mechanism between these policy maps, leading to final action expression.

2 The Five Factor Theories

The first of the five factor theories takes the form of *Stimulus-Response (S-R) Behaviorism*; which holds that action (the "response") selection is determined by the current sensory condition (the "stimulus"). Although first proposed in the final years of the 19th century [Thorndike, 1898], the approach continues to find contemporary support in the work of [Brooks, 1991; Bryson, 2000; and Maes, 1991]. This behavior is not defined by degree. The stimulus-response unit could be as apparently simple as a low-level reflex, such as the blink of an eye in response to a puff of air. Alternatively, behavioral repertories of considerable complexity can be postulated from essentially reactive models [Tyrrell, 1993; Tinbergen, 1951]. Such behaviors are generally considered to be innate (genetically determined) to the individual. Learning in the behaviorist regime is reward based, strengthening or weakening the connection between stimulus and response. It may be conjectured that not all such behaviors will be amenable to learning at the same rate, if at all.

The second factor theory, *classical conditioning*, was proposed by Ivan Pavlov (1849-1936) following observations that some innate reflexes can be associated with an otherwise neutral stimulus by repeated pairing, which will in turn elicit the reflex action. The procedure is highly repeatable and is easily demonstrated across a wide range of reflexes and species, and has been extensively modeled both mathematically and by implementation (e.g. [Vogel *et al*., 2004], for recent review).

The third theory, *operant* or *instrumental conditioning*, proposed by B.F. Skinner (1904-1990), who argued that actions were not "elicited" by impinging sensory conditions, but "emitted" by the animal in anticipation of a desired reward outcome. The effect is also highly repeatable under appropriate conditions, and it is clear that, given a suitable source of reward, an animal's (or indeed, a person's) behavior can be modified ("shaped") at will by judicious application of this principle. Whilst enormously influential in its time, only a relatively small number of computer models follow this approach (e.g. [Saksida *et al*., 1997], or Schmajuk [1994] implementing Mowrer's [1956] "two-factor" theory, incorporating both classical and operant conditioning effects.)

The fourth theory, the "cognitive" model, proposed by E.C. Tolman [1932] describes a three-part basic cognitive unit, which establishes the expectation or anticipation of a specific stimulus following, and contingent on, an action taken in the immediate context of another stimulus. The context stimulus and action provide the means to achieve a desired and anticipated stimulus, the end. Tolman's *means-ends* approach both inspired and continues to be a fundamental technique of problem solving and planning for artificial intelligence ([Russell and Norvig, 1995], for instance). The Dynamic Expectancy Model (DEM) [Witkowski, 1998; 2000; 2003] and the Anticipatory Classifier System (ACS) model [Stoltzmann *et al*., 2000] represent recent three-part cognitive models.

A fifth theoretical position, broadly characterized by the term *associationism* (e.g. [Hebb, 1949]), concerns the direct associability and anticipation of stimuli following repeated pairing of activations. While of greater significance in other aspects of animal modeling, this approach does not directly incorporate an action component, and discussion of it will be restricted here to a minor supporting role in the action selection problem.

3 Sense, Action and Valence

For largely historical reasons sensations are widely referred to as *stimuli* in this body of literature and the actions or behaviors generated as *responses*. This is not entirely satisfactory, as it largely fails to capture the range of interpretations required by the five theories taken together. Consequently, this paper will refer to the sense-derived component as a *sensory signature* or *Sign*, and denote such events by the symbol S, sub-scripts will be used to differentiate Signs were necessary. The philosophically neutral term *sense data* might also be employed for this purpose (e.g. [Austin, 1962]). In the SRS/E model, S := {0,1}.

Equally, the term "response" seems pejorative, and the more neutral term *Action* will be preferred, similarly abbreviated to A. Each Action will have associated with it an *action cost*, *ac*, (in SRS/E, by definition, $ac \geq 1$) indicating the time, effort or resource required to perform it.

Any Action may also be assigned an activation level, determined according to the rules presented later. Once activated, an Action becomes a candidate for *expression*, in which the Action is performed by the animal and may be observed or measured directly.

A Sign will be defined as a conjunction of detectable conditions (or their negations, acting as inhibitory conditions), typically drawn directly from the senses. Any Sign where all the conditions currently hold is said to be *active*. A Sign may be activated by some very specific set of detected sensory conditions, or be active under a wide range of conditions, corresponding to highly differentiated or generalized sensing.

Any Sign that is anticipated, but not active, is termed *sub-active*. Sub-activation is a distinct condition from full activation. It is important to distinguish the two, as the prediction of a Sign event is not equivalent to the actual event, and they have different propagation properties.

Additionally, any Sign may assume a level of *valence* (after [Tolman, 1932]), the extent to which that Sign has goal like properties, indicating that it may give the appearance of driving or motivating the animal to directed action selection behavior. Valence may be positive (goal seeking or rewarding) or negative (initiating avoidance

behaviors or being aversive). A greater valence value will be taken as more motivating, or rewarding, than a lesser one. Some Signs will hold valence directly, some via propagation to other Signs holding valence.

As with activation and sub-activation, the valence and sub-valence properties may also be propagated between Signs under the conditions described in section 5. A Sign that is the direct source of valence is deemed *satisfied* once it has become active, and it and the propagated chain of sub-valenced Signs will revert to their normal, unvalenced, state (unless there are multiple sources of direct valence).

4 The Forms of Connection

The anticipatory stance proposes that the principal effects of the five target theories can be adequately explained by adopting a combination of three connection types, and that their underlying function is to provide a temporally predictive link between different Sign and Action components. While noting that the model described here is highly abstracted, its biologically inspired background grounds it in the notion that, in nature, these abstract links represent physical neural connections between parts of the animal's nervous system and brain. These links, and such properties as sub-activation and valence, represent conjectures (from experimental observation) about the function of the brain that may be corroborated or refuted by further investigation.

With the exception of a connection of type **C1**, the abstract link types proposed below are bi-directional. Propagation effects across these links are asymmetric, and these properties are discussed in section 5.

This is not intended to imply that there are "bi-directional neurons", only that the structures that construct these linking elements have a complexity suited to the task. Where the animal does not possess a link or type of link (on the basis of its genetic makeup) it will be congenitally incapable of displaying a corresponding class of action selection behavior or learning. Of course, there are many other possible connection formats between arbitrary combinations of Signs and Actions; but it will be argued that these are sufficient to explain the principal properties of the five factor theorems.

Connection type C1 (SA): $S_1 \overset{w}{\rightarrow}_{t\pm\tau} (A \wedge S_2)$

Connection type C2 (SS): $S_1 \overset{v,c}{\leftrightarrows}_{t\pm\tau} S_2$

Connection type C3 (SAS): $(S_1 \wedge A) \overset{v,c}{\leftrightarrows}_{t\pm\tau} S_2$

While connections of type **C1** have only an implicit anticipatory role, connection types **C2** and **C3** are both to be interpreted as making explicit anticipatory predictions.

The type **C1** connection ("SA") is a rendition of the standard S-R behaviorist mechanism, with a forward only link from an antecedent sensory condition initiating (or at least predisposing the animal to initiate) the action A, as represented by the link "→". This symbol should definitely not be associated with logical implication, its interpretation is causal not truth preserving. The symbol *t* will indicate temporal delay (with range "±τ"), which may be introduced between the sense and action parts. The (optional) Sign S_2 is postulated as a mechanism for reinforcement learning, and is not required where learning across the connection (updating *w*) is not observed. The conjunctive connective symbol "∧" should be read as "co-incident with".

In keeping with standard behaviorist modeling, *w* will stand to indicate the strength, or *weight*, of the connection. This weight value will find application in selecting between candidate connections, and in considering reinforcement learning. Traditionally, the strength of the stimulus and a habituation mechanism for the action would also be postulated ([Hull, 1943], for a comprehensive discussion of these and related issues). Specifically the strength or likelihood of the response action will be modulated by the strength of the stimulus Sign.

4.1 Explicitly Anticipatory Connection Types

Connection type **C2** notates a link between two Signs, and indicates that Sign S_1 anticipates or predicts the occurrence of Sign S_2 within the specific time range $t\pm\tau$ in the future. This is indicated by the right facing arrow in the link symbol "⇆". The link has a *corroboration value*, *c*, associated with it, indicating the reliability of that prediction, based on continuing prior observation. A generic corroboration value update rule will be considered in section 6.1.

The *valence value*, *v*, of S_1 is a function of the current value of the valence value of S_2, and is hence associated with the left facing part of the link. Where the value $t\pm\tau$ is near zero, the link is essentially symmetric, S_1 predicts S_2 as much as S_2 predicts S_1. This is the classical Hebbian formulation. Where *t* is greater than zero (negative times have no interpretation in this context), the link is considered asymmetric. The assertion that S_1 predicts S_2 is no indicator that S_2 also predicts S_1. As the relationship between the two Signs is not necessarily causal, the animal may hold both hypotheses simultaneously and independently, as separate **C2** connections.

The **C3** connection differs from **C2** by the addition of an instrumental Action on the left hand side. The prediction of S_2 is now contingent on the simultaneous activation of both S_1 and the action A. The interpretation of the corroboration value *c* and the temporal offset *t* and range τ remain the same. The transfer of valence *v* to S_1 needs to now be a function of both S_2 and the action cost of A. This connection can be read as "the Sign S_2 is anticipated at time *t* in the future as a consequence of performing the action A in the context of S_1". Equally, it may serve as an instrumental operator: "to achieve S_2 at time *x* in the future, achieve S_1 at time *x-t*, and perform action A". Such links also take the form of independent hypotheses, giving rise to specific predictions that may be corroborated.

5 The Forms of Propagation

The five "rules of propagation" presented in this section encapsulate the operations on the three connection types with regard to the five factor theories. The rules define (i) when an Action becomes a candidate for expression, (ii)

when a Sign will become sub-activated, (iii) when a prediction will be made, and (iv) when a Sign will become valenced by propagation.

In the semi-formal notation adopted below active(), sub_active(), expressed(), valenced() and sub_valenced() may be treated as predicate tests on the appropriate property of the Sign or Action. Thus active(S_1) will be asserted if the Sign denoted by S_1 is active. The disjunction "∨" should be read conventionally as either or both, the conjunction "∧" should be interpreted as in section 4. On the right hand side of the rule, activate(), predict() and sub_valence() should be taken as "internal actions", operations taken to change the state or status of the item(s) indicated.

Rule P1 Direct Activation:
For any **C1** (SA) link,
if (active(S_1) ∨ sub_active(S_1))
then activate(A, *w*)

Rule P2 Sign Anticipation:
For any **C2** (SS) link,
if (active(S_1) ∨ sub_active(S_1))
then sub_active(S_2)

Rule P3 Prediction:
For any **C2** (SS) link,
if(active(S_1))
then predict(S_2, $t \pm \tau$)
For any **C3** (SAS) link,
if(active(S_1) ∧ expressed(A))
then predict(S_2, $t \pm \tau$)

Rule P4 Valence transfer:
For any **C2** (SS) link,
if(valenced(S_2) ∨ sub_valenced(S_2))
then sub_valence(S_1, $f(v(S_2), d)$)
For any **C3** (SAS) link,
if(valenced(S_2) ∨ sub_valenced(S_2))
then sub_valence(S_1, $f(v(S_2), c, ac(A))$)

Rule P5 Valenced activation:
For any **C3** (SAS) link,
if(active(S_1) ∧ sub_valenced(S_1))
then activate(A, *v'*)

Rule **P1** expresses the standard S-R behaviorist rule. Only in the simplest of animals would the activation of the action A lead to the direct overt expression of the action or activity. As there is no assumption that Signs are mutually exclusive, many actions may become candidates for expression. The simplest strategy involves selecting a "winner" based on the weightings and putting that action forward to the effector system for external expression.

Rule **P2** allows for the propagation of sub-activation. The effect is instantaneous, notifying and allowing the animal to modify its action selection strategy immediately in anticipation of a possible future event. Evidence from second order classical conditioning studies would suggest that sub-activation propagates poorly (i.e. is heavily discounted).

Rule **P3** allows for a specific prediction of a future event to be recorded. This calls for a limited form of memory of possible future events, analogous to the more conventional notion of a memory of past events. Under this formulation, predictions are created as a result of full activation of the Sign and actual expression of the Action, and are therefore non-propagating. Predictions are made in response to direct sense and action and are employed in the corroboration process (section 6.1). This process is distinct from sub-activation, which is propagating, but non-corroborating.

Rule **P4** indicates the spread of valence backwards along chains of anticipatory links. The sub_valence() process is shown in different forms for the **C2** (SS) and **C3** (SAS) links, reflecting the discounting (*d*) process mentioned earlier. As an exemplar, in the SRS/E model valence is transferred from S_2 to S_1 across the **C3** link according to the generic formulation: $v(S_1) := v(S_2) * (c / ac(A))$. By learning rule **L2** and **L3** (section 6.1) $0 < c < 1$, and as $ac(A) \geq 1.0$ (by definition), therefore $v(S_1) < v(S_2)$. Valence propagates preferentially across high confidence links with "easy" (i.e. lower cost value) Actions. Transfer is straightforward and has proved robust in operation in the DEM and SRS/E.

Rule **P5** indicates the activation of any Action A where the antecedent Sign S_1 is both active and valenced. As with rule **P1**, many Actions may be affected. The one associated with the highest overall S_1 valence value is selected.

The choice process by which the various activated Actions give rise to the action to be selected for overt expression is the subject of section 8. For a simple S-R only (rule **P1**) system, this might be summarized as selecting the action associated with the highest weight value, but there must be a balance between the actions activated by rule **P1** and those by **P5**. Note here that the valence value *v'* refers to the valence value of the Sign holding direct valence (the *top-goal*), whose value has been propagated to the SAS link, not that of either S_1 or S_2 of the **C3** (SAS) link in question.

6 The Forms of Learning

This section describes the conditions under which learning will take place. In the anticipatory action selection model presented, the net effect of learning is to modify the Actions or activities to be expressed (and so the observable behavior of the animal) in response to a particular Sign. Each of the five factor theories takes a particular stance on the nature of learning.

In the first, *reward based learning*, learning is taken to be a consequence of the animal encountering a valenced situation following an action – one that is characterised as advantageous/disadvantageous and thus interpreted as "rewarding" (or not) to the animal. This is frequently referred to as reinforcement learning. There are a wide range of reinforcement learning methods, so a generic approach will be adopted here.

In the second, *anticipatory learning*, "reward" is derived from the success or otherwise of the individual predictions made by the propagation rules given in section 5. In one sense, the use of link type **C3**, as described here, can be seen as subsuming link type **C1**, but the converse does not

hold. In the **C1** link, the role of anticipation in the learning process is implicit but is made explicit in the **C3** type link.

Learning rule L1 (the reinforcement rule):
For any **C1** (SA) link
if (active(A) ∧ (valence(S_2) ∨ sub_valence(S_2)))
then update(w, α)

This is a generic form of the standard reinforcement rule. If the action is associated with any sensation that provides valence, then the connection weight w will be updated asymptotically by some factor α. Several well established weight update strategies are available, such as Watkins' *Q-learning* and Sutton's *temporal differences* (TD) method, see [Sutton and Barto, 1998] for review. In each the net effect is to increase or decrease the likelihood that the link in question will be selected for expression in the future.

6.1 Methods of Anticipatory Learning

A central tenet of the anticipatory stance described in this paper is that certain connective links in the model make explicit predictions when activated. Recall that propagation rule **P3** creates explicit predictions about specific, detectable, events that are anticipated to occur in the future, within a specific range of times (denoted by t±τ). The ability to form predictions has a profound impact on the animal's choice for learning strategies. This section considers the role played by the ability to make those predictions.

Learning rule L2 (anticipatory corroboration):
For any (**C2** ∨ **C3**) link
if(predicted(S_2, -t±τ) ∧ active(S_2))
then update(c, α)

Learning rule L3 (anticipatory dis-corroboration):
For any (**C2** ∨ **C3**) link,
if(predicted(S_2, -t±τ) ∧ ¬active(S_2))
then update(c, β)

Learning rule L4 (anticipatory link formation):
if(¬predicted(S_x)),
then create_SAS_link(S_y, A_y, S_x, t, τ)
or create_SS_link(S_y, S_x, t, τ)

These three rules encapsulate the principles of anticipatory learning, and are applicable to both **C2** and **C3** link types. Three conditions are significant, where a prediction has been made, and the predicted event did occur at the expected time (learning rule **L2**). The link is considered corroborated and is strengthened. Where a prediction is made, but the event does not occur (learning rule **L3**), the link is considered dis-corroborated and weakened. Lastly, where an event occurs, but it was not predicted at all (learning rule **L4**).

The SRS/E computer implementation employs the simple but robust, effective and ubiquitous update rule $c := c+\alpha\,(1-c)$, where $(0 \le \alpha \le 1)$ for **L2**, and the generic update rule $c := c-\beta(c)$, where $(0 \le \beta \le 1)$, is again simple, effective and robust for **L3**. Both update functions are asymptotic towards 1.0 and zero respectively. The net effect of these update rules is to maintain a form of "running average" more strongly reflecting recent outcomes, with older outcomes becoming successively discounted (tending to zero contribution). The greater the values of α and β, the more aggressively recent events are tracked. The particular settings of these values are specific to the individual animal. Where no prediction was made by a rule, c remains unchanged regardless of the occurrence of S_2. This is consistent with the notion that a rule is only responsible for predicting an event under the exact conditions it defines.

The key issue here is that anticipatory learning is *everytime*. Every prediction made, regardless of its cause, initiates learning. Learning is independent of valenced reward (this is the phenomenon of *latent learning* [Witkowski, 1998], [Thistlethwaite, 1951]). Anticipatory links are measured relative to their predictive ability, not their usefulness. Correct anticipation is its own reward. Such anticipatory reward is generated locally to the **C2** or **C3** link, and is independent of all others. Further, if circumstances change, each link adjusts automatically to the prevailing circumstances based on recent predictive experience. Anticipation may also be combined with valence, to preferentially focus the learning process on Signs that have, or have had, valence (e.g. the *Valence Level Pre-Bias* technique [Witkowski, 1998]).

Where an event is unpredicted by any link, this is taken as a cue to establish a new link between the unpredicted event (as S_2) and some recent recently active event (as S_1) at time t, rule **L4**. Where a **C3** link is created some expressed Action A contemporary with the new S_1 is also implicated. Again the choice of how many new links are formed, and the range of values for t and τ are specific to the individual animal. Without any *a-priori* indication as to which new links might be effective, higher learning rates can be achieved by forming many links, and then allowing learning rules **L2** and **L3** separate the effective from the ineffective.

The key issue here is that link learning may be invoked everytime a novel or unpredicted Sign is detected. Learning may proceed from *tabula rasa*, and is rapid while much is novel. In a restricted environment, link learning will slow as more is predicted, but resume if circumstances change.

No rule for link removal is considered here, but has been discussed elsewhere in the context of the DEM. Witkowski [2000] considers the rationale for retaining links even when their corroboration values fall to very low values, based on evidence from behavioral extinction experiments [Blackman, 1974].

7 Explaining the Five Factors

This section returns to the action selection factor theories outlined in section 2, and will discuss them in turn in terms of the link types, propagation rules and learning rules presented and discussed in sections 4, 5 and 6. As previously indicated, each theory supports and is supported by an (often substantial) body of experimental evidence, but that each theory in turn fails to capture and explain the overall range of action selection behaviors displayed by any particular animal or species. The conceptually simpler

approaches are covered by single links and rules, others require a combination of forms, and yet others perhaps require re-interpretation in the light of this formulation.

7.1 Stimulus-Response Behaviorism

S-R Behaviorism holds that all, or the majority of, observed and intelligent behavior can be ascribed to an innate, pre-programmed, pairing of sense data driven stimuli and pre-defined actions.

7.1.1 Static Policy Maps

With no embellishments, S-R behaviorism is reduced to connection type **C1** and propagation type **P1**. The underlying assumption that all these strategies adopt is to tailor the behavior of the organism, such that the actions at one point sufficiently change the organism or its environment such that the next stage in any complex sequence of actions becomes indicated. We may refer to this as a *static policy map*. The DEM records these connections in a list, effectively ordered by the weight parameter, *w*. Recall that the weighting value *w* may be modified by reinforcement learning [Sutton and Barto, 1998].

Given a sufficient set of these reactive behaviors, the overall effect can be to generate exceptionally robust behavioral strategies, apparently goal seeking, in that the actually independent elements of sense, action and actual outcome combinations, inexorably leads to food, or water, or shelter, or a mate [Bryson, 2000; Maes, 1991; Tinbergen, 1951; Tyrrell, 1993].

Such strategies can appear remarkably persistent, and when unsuccessful, persistently inept. Any apparent anticipatory ability in a fixed S-R strategy is not on the part of the individual, but rather a property of the species as a whole. With sufficient natural diversity in this group strategy, it can be robust against moderate changes in the environment, at the expense of any individuals not suited to the changed conditions.

7.2 Classical Conditioning

Reactive behaviorism relies only on the direct activity of the Sign S_1 to activate A, this is the *unconditioned response* (UR) to the *unconditioned stimulus* (US): the innate reflex. As reflexes are typically unconditionally expressed (i.e. have high values of *w*) the US invariably evokes the UR. Rule **P1** allows for sub-activation of the S_1 Sign. Therefore, if an anticipatory **C2** connection is established between a Sign, say S_X and the US Sign S_1, then activation of S_X will sub-activate S_1, and in turn evoke A, the *conditioned response* (CR).

Note the anticipatory nature of the CS/US pairing [Barto and Sutton, 1982], where the CS must precede the US by a short delay (typically <1s). The degree to which the CS will evoke CR depends on the history of anticipatory pairings of S_X and S_1, and is dynamic according to that history, by learning rules **L2** and **L3**, the rates depending on the function of α and β. If the link between CS and US is to be created dynamically, then learning rule **L4** is invoked. The *higher order conditioning* procedure allows a second neutral Sign (S_Y) to be conditioned to the existing CS (S_X), using the standard procedure: S_Y now evokes the CR. This is as indicated by the propagation of sub-activation in **P2**.

Overall, the classical condition reflex has little impact on the functioning of the policy map of which its reflex is a part. Indeed the conditioned reflex, while widespread and undeniable, could be thought of as something of a curiosity in learning terms (B.F. Skinner reportedly held this view). However, it provides direct, if not unequivocal, evidence for several of the rule types presented in this paper.

7.3 Operant Conditioning

Operant conditioning shapes the overt behavior of an animal by pairing the actions it takes to the delivery of reward. The experimenter need only wait for the desired action and then present the reward directly. This is typified by the *Skinner box* apparatus, in which the subject animal (typically a hungry rat) is trained to press a lever to obtain delivery of a food pellet reward. We interpret this link as an anticipatory one. The action anticipates the sensory condition (food), which, as the rat is hungry, holds valence. Further, the experimenter might present the food only when the action is taken in some circumstances, not others. The animal's behavior becomes *shaped* to those particular circumstances. These are the conditions for the **C3** connection type. This is equivalent to Catania's [1988] notion of an operant *three-part contingency* of "stimulus - response - consequence".

The association between lever (S_1), pressing (A) and food (S_2) is established as a **C3** (SAS) link by **L4**. When the action is preformed in anticipation of S_2, the link is maintained, or not, by **L2** and **L3** according to the outcome of the prediction made (**P3**). While food (S_2) retains valence, and the rat is at the lever, the rat will press the lever (**P5**), and in the absence of any alternative, continue to do so. Action selection is now firmly contingent on both encountered Sign and prevailing valence.

Due to valence transfer (**P4**) such contingencies propagate. Were the rat to be in the box, but not at the lever, and some movement A_M would take to rat from its current location S_C to the lever S_L, then the **C3** contingency $(S_C \wedge A_M) \leftrightarrows S_L$ would provide propagated valence to S_C and result in A_M being activated for expression. Once the rat is satiated, the propagation of valence collapses and the expression of these behaviors will cease. This transfer of valence may be used to create long chains of behaviors (such as in preparing animals for film performances) by building the sequence back one step at a time.

Propagation rule **P4** also allows for *secondary* or *derived reinforcement* effects ([Bower and Hilgard, 1981], p.184), in which normally non-reinforcing **C2** links may be paired with (or even chained from) an innately valenced one.

7.4 Tolman's Expectancy Model

Catania's [1988] description of the operant three-part contingency, described in the light of this formulation looks suspiciously like Tolman's [1932] *Sign-Gestalt Expectancy*, an explicitly anticipatory three-part Sign-Action-Sign (i.e. **C3**) link. Skinner, as a staunch "old-school" behaviorist, would definitely not have approved! Where the Skinner box

investigates the properties of the individual **C3** link, which may be explored in detail under a variety of different schedules, Tolman's work primarily used mazes. Rats, in particular, learn mazes easily, recognize locations readily and are soon motivated to run mazes to food or water when hungry or thirsty. Mazes are also convenient experimentally, as they may be created with any desired pattern or complexity.

Choice points and other locations in the maze may be represented as Signs (a rat may only be in one location at once, though it may be mistaken as to which one), and traversal between them as identifiable Actions. Every location-move-location transition may be represented as an anticipatory **C3** connection. Recall that these links are only hypotheses - errors, or imposed changes to the maze are accommodated by the learning rules **L2**, **L3** and **L4**.

It is now easy to see that, placed in a maze, the animal will learn the structure as a number of **C3** connections with or without (i.e. latently) the presence of valence or reward. Novel locations encountered by random (or guided) exploration invoke **L4**, and the confidence value c is updated each time a location is revisited, by **L2** or **L3**. Once encountered, food may impart valence to a location (by **P4**).

7.4.1 Dynamic Policy Maps

If at any time a location becomes directly or indirectly linked to a source of valence (i.e. food to a hungry rat), this valence will propagate across all the **C3** (and indeed **C2**) links to establish a *dynamic policy map* (DPM). This takes the form of a graph of all reachable Signs. In SRS/E this is considered as form of a modified breadth first search, in which each Sign node is assigned the highest propagated valence value. Again this generic process, as implemented in SRS/E, is computationally fast and robust in operation.

Once created, each Sign implicated in the DPM is associated with a single Action from the appropriate **C3** link, the one on the highest value valence path, and a single valence value v, indicating its rank in the policy map. Given this one to one, ordered mapping, an action may be selected from the DPM in a manner exactly analogous to a static policy map. In this respect, the behavior chaining technique described in section 7.3 looks to be no more than an attempt to manipulate the naturally constructed dynamic policy to prefer one chain of actions to all the others.

The dynamic policy map must be recomputed each time there is a change in valence or any learning event takes place (i.e. almost everytime). Sometimes this has little effect on the observable behavior, but sometimes has a dramatic and immediate effect, with the animal reversing its path or adopting some completely new activity.

Figure 1 illustrates this (from [Witkowski, 2000]). The animat (circle) is in a grid maze, and each square represents a location Sign, and the arrows indicate the current policy action in that square. The animat was allowed to explore the maze shown on the left completely by selecting random actions, but without any source of valence (i.e. latently). When G is given valence, the animat builds a DPM and takes the shortest path via B. With the animat returned to S, and G still valenced, but B now blocked, the DPM will still indicate a path via B (the blockage is undiscovered), center. As the intended (up) action to B now fails, the DPM alters to prefer the apparently longer path via A, and the observable behavior of the animat will abruptly change as a new DPM is constructed and a new path is preferred, right.

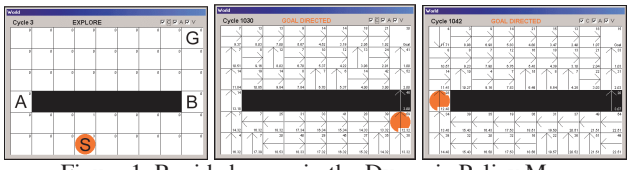


Figure 1: Rapid changes in the Dynamic Policy Map

8 Combining Static and Dynamic Policy Maps

For any animal that displays all the forms of action selection, it becomes essential to integrate the effects of innate behaviour, the static policy map, and the valence driven dynamic policy map. The dynamic policy map is transient, and must interleave with the largely permanent static policy map. The valence value of the original source (v' from section 5, the top-goal) is (numerically) equated to the **C1** connection weight values, w. While $v' \geq \text{active}(w)$, actions are selected only from the DPM. If at any point $\text{active}(w) > v'$, DPM selection is suspended, and actions are taken from the static policy. Once completed or abandoned, control reverts to the DPM.

This allows for high-priority activities, such as predator avoidance, to invariably take precedence over goal-seeking activities. As the valence of the goal task increases, the chance of it being interrupted in this way decreases. After an interruption from static actions, valenced action selection resumes, The DPM must be reconstructed, as the animal's situation will have been changed, and the static actions may also have given rise to learned changes – a case of everytime learning.

Static policy maps may also be partitioned. Tinburgen [1951] proposed the use of hierarchical *Innate Releaser Mechanisms* (IRM) to achieve this. In each case, the releasing enabler should take its place in the static ranking, with all its subsidiary SR connections simultaneously enabled, but then individually ranked within that grouping. Selection may then proceed as for the Dynamic Policy Map example. Note that in the DEM, valence setting is reserved as a static policy map activity, a type of Action. In this context the IRM releasing enablers start to look, in evolutionary terms, like the beginnings of valenced items.

9 Summary and Conclusions

This paper has presented a high-level view of the action selection properties of five central theories of behavior and learning. Each of these theories holds that actions are selected on the basis of prevailing sensory conditions. They do not agree on how this occurs, yet it is clear that it may be demonstrated experimentally that each theory accounts for only a part of an individual animal's behavioral repertoire, and that what the experimenter sees is at least partly due to the design of their experiments. The paper has developed a

set of five propagation rules and four learning strategies over three connection types to encapsulate and unify these otherwise apparently disparate approaches.

This has lead to the notion of different types of policy map operating within the animal, from static to dynamic, and how they may be combined to exhibit apparently different behavioral phenomena under the variety of circumstances the animal may encounter, in nature or the laboratory. The Dynamic Expectancy Model has been employed as an implemented (SRS/E) framework for this discussion.

Much remains to be done. This overview paper has laid a ground plan, but the devil remains in the detail. There exists a truly vast back catalogue of experimental data from the last 100 years of investigations that might be revisited in the light of this framework. Two substantive questions remain: (i) whether the links, propagation and learning rules presented sufficiently describe the five factor theories, and (ii) whether, even taken together as a whole, the five factor theories are sufficient to explain all animal behavior.

On the first, the theories are based on these experiments, and much falls into place as a consequence. On the second, it seems unlikely - as evolutionary pressure has lead to incredibly diverse behavior patterns and mechanisms. Identifying these experimentally observed exceptions will serve to refine the multi-factor approach presented, leading in time to a better, more encompassing, solution.

Even though one can observe classical and operant conditioning, and means-ends behavior in humans, it is abundantly clear than even taken together the five factors fail to explain human behavior to a very considerable extent. It is vastly apparent that human (and possibly other primate) activities are not solely, or even predominantly, driven directly by immediately prevailing and observable circumstances. However, one might see these five mechanisms as both a foundation for, and a bridge to, the evolutionary development of higher-level cognitive functions.

References

[Austin, 1962] Austin, J.L. *Sense and Sensibilia*, Oxford University Press, 1962

[Barto and Sutton, 1982] Barto, A.G. and Sutton, R.S. Simulation of Anticipatory Responses in Classical Conditioning by a Neuron-like Adaptive Element, *Behavioral Brain Research*, **4**:221-235, 1982

[Blackman, 1974] Blackman, D. *Operant Conditioning: An Experimental Analysis of Behaviour*, London: Methuen & Co.

[Bower and Hilgard, 1981] Bower, G.H. and Hilgard, E.R. *Theories of Learning*, Englewood Cliffs: Prentice Hall Inc., fifth edition, 1981

[Brooks, 1991] Brooks, R.A. Intelligence Without Reason, *MIT AI Laboratory, A.I. Memo No. 1293*. (Prepared for Computers and Thought, IJCAI-91, pre-print), April, 1991

[Bryson, 2000] Bryson, J. Hierarchy and Sequence vs. Full Parallelism in Action Selection, *6th Int. Conf. on Simulation of Adaptive Behavior (SAB-6)*, pages 147-156, 2000

[Catania, 1988] Catania, A.C. The Operant Behaviorism of B.F. Skinner, in: Catania, A.C. and Harnad, S. (eds.) *The Selection of Behavior*, Cambridge University Press, pages 3-8, 1988

[Hebb, 1949] Hebb, D.O. *The Organization of Behavior*, John Wiley & Sons, 1949

[Hull, 1943] Hull, C. *Principles of Behavior*, New York: Apple-Century-Crofts, 1943

[Maes, 1991] Maes, P. A Bottom-up Mechanism for Behavior Selection in an Artificial Creature, *1st Int. Conf. on Simulation of Adaptive Behavior (SAB)*, pages 238-246, 1991

[Mowrer, 1956] Mowrer, O.H. Two-factor Learning Theory Reconsidered, with Special Reference to Secondary Reinforcement and the Concept of Habit, *Psychological Review*, **63**:114-128, 1956

[Russell and Norvig, 1995] Russell, S. and Norvig, P. *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.

[Saksida *et al*., 1997] Saksida, L.M., Raymond, S.M. and Touretzky, D.S. Shaping Robot Behavior Using Principles from Instrumental Conditioning, *Robotics and Autonomous Systems*, **22**-3/4:231-249, 1997

[Schmajuk, 1994] Schmajuk, N.A. Behavioral Dynamics of Escape and Avoidance: A Neural Network Approach, *3rd Int. Conf. on Simulation of Adaptive Behavior (SAB-3)*, pages 118-127, 1994

[Stoltzmann *et al*., 2000] Stolzmann, W., Butz, M.V., Hoffmann, J. and Goldberg, D.E. First Cognitive Capabilities in the Anticipatory Classifier System, *6th Int. Conf. on Simulation of Adaptive Behavior (SAB-6)*, pages 287-296, 2000

[Sutton and Barto, 1998] Sutton, R.S. and Barto, A.G. *Reinforcement Learning: An Introduction*, Cambridge, MA: MIT Press, 1998

[Thistlethwaite, 1951] Thistlethwaite, D. A Critical Review of Latent Learning and Related Experiments, *Psychological Bulletin*, **48**-2:97-129, 1951

[Tinbergen, 1951] Tinbergen, N. *The Study of Instinct*, Oxford: Clarendon Press, 1951

[Thorndike, 1898] Thorndike, E.L. Animal Intelligence: An Experimental Study of the Associative Processes in Animals, *Psychol. Rev., Monogr. Suppl*., **2**-8, 1898

[Tolman, 1932] Tolman, E.C. *Purposive Behavior in Animals and Men*, New York: The Century Co., 1932

[Tyrrell, 1993] Tyrrell, T. *Computational Mechanisms for Action Selection*, University of Edinburgh, Ph.D. thesis, 1993

[Vogel *et al*., 2004] Vogel, E.H., Castro, M.E. and Saavedra, M.A. Quantitative Models of Pavlovian Conditioning, *Brain Research Bulletin*, **63**:173-202, 2004

[Witkowski, 1998] Witkowski, M. Dynamic Expectancy: An Approach to Behaviour Shaping Using a New Method of Reinforcement Learning, *6th Int. Symp. on Intelligent Robotic Systems*, pages 73-81, 1998

[Witkowski, 2000] Witkowski, M. The Role of Behavioral Extinction in Animat Action Selection, *proc. 6th Int. Conf. on Simulation of Adaptive Behavior (SAB-6)*, pages 177-186, 2000

[Witkowski, 2003] Witkowski, M. Towards a Four Factor Theory of Anticipatory Learning, in Butz, M.V. *et al*. (Eds.) Anticipatory Behavior in Adaptive Learning Systems, Springer LNAI 2684, pages 66-85, 2003

On Compromise Strategies for Action Selection with Proscriptive goals

Frederick Crabbe
U.S. Naval Academy
Computer Science Department
572M Holloway Rd, Stop 9F
Annapolis, MD 21402
crabbe@usna.edu

Abstract

Among many properties suggested for action selection mechanisms, one prominent one is the ability to select compromise actions, i.e. actions that are not the best to satisfy any active goal in isolation, but rather compromise between the multiple goals. This paper performs an analysis of compromise actions in situations where the agent has one proscriptive goal. It concludes that optimal compromise behavior looks quite different from what was expected, and, while optimal compromise actions are beneficial to an agent, the benefit is often small compared to greedy algorithms. It goes on to suggest that much of the discussion about compromise behavior is the result of an equivocation on its definition, and it proposes a new compromise behavior hypothesis.

1 Introduction

Traditional Artificial Intelligence planning systems use search in order to fully characterize the space of actions a robotic agent can select in a given situation. The agent considers the outcomes of possible actions into the future until it finds sequences of actions that achieve its goals. One feature of this approach is that given enough time, a planning system can determine the optimal action sequence for the agent. Of course, the issue of time is a fundamental problem for these planning systems: the agent may not have at its disposal the time needed in order to discover the optimal actions—in fact, often the amount of time required exceeds the age of the universe.

Behavior-based approaches to robotics and agents in general have been introduced to address these sorts of problems [Brooks, 1986; Arkin, 1998]. These distributed reactive-style approaches are designed to generate "good enough" actions in a very small amount of time. Without optimality, there arises the important question of exactly what "good enough" means. In his now classic Ph.D. thesis, Tyrrell introduced a list of fourteen requirements for Action Selection Mechanisms. Of these, number twelve was "Compromise Candidates: the need to be able to choose actions that, while not the best choice for any one sub-problem alone, are best when all sub-problems are considered simultaneously." [Tyrrell, 1993, p. 174] Tyrrell's list has had significant impact on the Action Selection field [Humphrys, 1996; Decugis and Ferber, 1998; Bryson, 2000; Girard *et al*., 2002, e.g.], and a number of researchers have developed systems to meet the criteria he set out [Werner, 1994; Blumberg, 1994; Crabbe and Dyer, 1999; Avila-Garcia and Canamero, 2004, e.g.]. Meanwhile biologists and ethologists have noted apparent compromise among animals in several scenarios.

The ability to consider compromise actions in an uncertain world makes great intuitive sense. When multiple goals interact, solving each optimally is not always optimal for the overall system. Yet, recent work has generated empirical results that seem to contradict the claim that the ability to consider compromise candidates is necessary [Jones *et al*., 1999; Bryson, 2000; Crabbe, 2004]. Despite this, there have been few in-depth analyses of the nature of compromise actions and their effect on the overall success of an agent. This paper presents an extension of the work by Hutchinson [1999] and Crabbe [2004] to investigate the nature of compromise actions in various environmental conditions, concluding that: optimal compromise behavior is qualitatively different from what might be expected; optimal compromise behavior provides less benefit than expected in the scenarios tested; and the apparent disagreement about the utility of compromise behavior might possibly arise from an equivocation on its definition.

2 Problem Formulation

The action selection problem we will discuss in this paper depends on the types of actions the agent can select, the types of goals the agent pursues, and the formal representation of the problem.

2.1 Actions

When designing an action selection system, the character of the "actions" selected by the agent affect the behavior exhibited. For instance, there is a clear difference between an agent selecting the action *contract left quadricep 3 cm*. and the action *go to the refrigerator*. The distinction is based on the level of specificity given by the action; the former is as specific as possible, while the latter leaves much room for interpretation on how it is to be accomplished. In this paper we will define our domain to be that of navigation of a mobile agent, similar to several authors' simulated domains [Maes, 1990; Tyrrell, 1993] or navigating mobile robots [Choset *et al*., 2005]. The space will be continuous, but time will be discrete, such that the action at each time step is defined as a movement 1 distance unit at any angle. The importance of this choice of the definition of "action" will be discussed in Section 6.

2.2 Goals

Tyrrell famously defined compromise as follows: "a *compromise candidate*, which might be beneficial to two or more systems to an intermediate degree, may be preferable to any of the candidates which are most beneficial for one system alone." [1993, p. 170] The problem of compromise in action selection has multiple guises. One fundamental distinction pivots on the nature of the involved goals: are they prescriptive or proscriptive? Prescriptive goals encourage an agent to take some action or sequence of actions in order to be satisfied. These goals are typically satisfied by a final consumatory act. Proscriptive goals encourage an agent to *not* perform certain actions in certain situations. These goals are typically not satisfied by a particular action, but can be said to have been satisfied over a period of time if offending actions are not performed. The nature of a possible compromise scenario changes depending on whether there are two prescriptive goals or one prescriptive and one proscriptive goal.

Two Prescriptive Goals

In a two-prescriptive-goal case, an agent has goals to be co-located with one of two target locations in the environment. These could be, for instance, the locations of food, water, potential mates, or shelter. At any moment either or both of the targets can disappear from the environment. The agent must select an action that maximizes its chances of co-locating with a target before it disappears. This model is drawn from several scenarios in biology. For example, frogs or cricket males sometimes advertise for mates by emitting calls. The males may disappear with respect to the female through a cessation of signaling. This can occur either due to the actions of predators, the arrival of a competing female, or for internal reasons such as energy conservation. Another scenario in biology is that of a hunter such as a cat stalking prey such as birds in a flock, where an individual bird can fly at any moment [Hutchinson, 1999]. Several action selection mechanisms, such as Werner [1994] and Montes-Gonzales et al. [2000] have been specifically designed to exhibit this sort of compromise. Further, biologists and ethologists have advocated in favor of the prescriptive version for some time [Morris *et al*., 1978; Lorenz, 1981; Latimer and Sippel, 1987; Bailey *et al*., 1990].

Crabbe [2004] gave strong evidence that while this sort of behavior is seen in nature, it confers little absolute advantage to the agent. In particular: the optimal compromise strategy performed only slightly better than the best non-compromise (or greedy) strategy and all other known compromise strategies perform worse than maximum expected utility. They conclude that "animals that exhibit apparent compromise [in the 2 prescriptive goal case] are either using some unknown strategy or are doing so for some other reason." [Crabbe, 2004] This paper discusses the implications and a possible explanation of Crabbe's result in greater detail in Section 6 below.

One Prescriptive, One Proscriptive Goal

Although the two-prescriptive-goal scenario has had significant impact on the action selection community, the more famous of the two compromise scenarios discussed here is when there is one prescriptive goal and one proscriptive goal.

> "...proscriptive sub-problems such as avoiding hazards should place a demand on the animal's actions that it does not approach the hazard, rather than positively prescribing any particular action. It is obviously preferable to combine this demand with a preference to head toward food, if the two don't clash, rather than to head diametrically away from the hazard because the only system being considered is that of avoid hazard" [Tyrrell, 1993]

As the quote above indicates, the idea that compromise actions are especially beneficial in the proscriptive goal case is intuitively appealing. Further, examples of this appear in the ethological literature. Blue herons will select sub-optimal feeding patches to avoid predation by hawks in years when the hawk attacks are frequent [Caldwell, 1986]. Similar behavior has been shown in sparrows [Grubb and Greenwald, 1982], minnows [Fraser and Cerri, 1982], pike and sticklebacks [Milinksi, 1986]. At the motor level, geese and other *anatidae* who are offered food by a human can sometimes exhibit behavior where the neck muscles for both a feeding behavior and a recoiling behavior are activated, causing a trembling in the neck [Lorenz, 1981].

The purpose of this paper is to provide an analysis of the proscriptive goal scenario using the techniques developed by Crabbe for the prescriptive goal scenario, to determine both the amount of benefit of compromise actions, as well as under what conditions compromise actions are the most useful.

2.3 Formal Model

To approximate the scenario described in the quote above, we examine a continuous environment with a target t and a danger d, corresponding to a resource such as a mate and a predator respectively. At any time the target can disappear from the environment (e.g. the prospective mate stops signaling) with a probability $1 - p_t$, and the danger can disappear (e.g. the predator becomes bored and wanders off) with a probability $1 - p_d$. That is, at each time step, the target remains in the environment with probability p_t and the danger remains in the environment with probability p_d. Also at each time step, there is a probability $p_n(d)$ that the predator will *not* strike or pounce on the agent. This probability is a function of the distance between the agent and the danger. The experiments in this paper use four different functions to generate the $p_n(d)$. The agent also has a goal level associated with the target and the danger, (G_t and G_d) that can vary with the quality of the resource and the damage due to the predator. Notationally, $\overline{i,j}$ is the distance from some location i to some location j. All distances are measured in the number of time steps it takes the agent to travel that distance.

3 Analytical Set-up

In order to investigate compromise candidates, we will analyze the initial configuration using Utility Theory [Howard, 1977]. Utility Theory assigns a set of numerical values (utilities) to states of the world. These utilities represent the usefulness of that state to an agent. Expected Utility (EU) is a prediction of the eventual total utility an agent will receive if it takes a particular action in a particular state. The Expected Utility (EU) of taking an action A in a state S is the sum of the product of the probability of each outcome that could occur and the utility of that outcome:

$$EU(A|S) = \sum_{S_o \in Outcomes} P(S_o|A,S)U_h(S_o) \qquad (1)$$

where $P(S_o|A,S)$ is the probability of outcome S_o occurring given that the agent takes action A in state S, and $U_h(S_o)$ is the historical utility of outcome S_o as defined below.

Assuming the agent is rational, the set of goals to consume objects will be order isomorphic[1] to the set of the agent's utilities of having consumed the objects. Therefore, EU calculated with utilities is order isomorphic with EU calculated with goals instead. For our purposes, we will assume that the goals and utilities are equivalent ($U(t) = G_t$).

Because a rational agent is expected to select the action with the largest EU, the historical utility of a state is the utility of the state plus future utility, or the max of the expected utility of the actions possible in each state:

$$U_h(S) = U(S) + \max_{A \in Actions} EU(A|S). \quad (2)$$

An agent can calculate EU using multiple actions in the future by recursively applying equations (1) and (2).

3.1 Optimal Behavior

We analyze compromise by comparing a close approximation of optimal behavior with several non-optimal but easy to generate behaviors. We approximate the optimal behavior based on the dynamic programming technique adapted by Crabbe from Hutchinson [1999]. This technique overlays a grid of points on top of the problem space and calculates the maximal expected utility of each location given optimal future actions. This is done recursively starting at the target locations and moving outward until stable values have been generated for all grid points.

The value we are trying to calculate is the expected utility of acting optimally at some location λ in a state where the target and the danger are still in the environment: $EU(O|t, d, \lambda)$. If θ is the angle of the optimal move for the agent at location λ and λ' is 1 unit away from λ in direction θ, then by equations 1 and 2 the expected utility of being at λ is:

$$\begin{aligned} EU(O|t, d, \lambda) =& p_t p_d p_n(\lambda) EU(O|t, d, \lambda') + \\ & p_t(1 - p_d) EU(O|t, \lambda') + \\ & p_d(1 - p_n(\lambda)) G_d, \\ & (1 - p_t) p_d p_n(\lambda) EU(O|d, \lambda') \quad (3) \end{aligned}$$

$$EU(O|t, \lambda) = G_t p^{\overline{\lambda,t}}, \text{ and,} \quad (4)$$

$$\begin{aligned} EU(O|d, \lambda) =& p_n(\lambda') p_d EU(O|d, \lambda') + \\ & (1 - p_n(\lambda')) G_d. \quad (5) \end{aligned}$$

The total expected utility is the expectation over four possible situations: both target and danger are still there, but the danger does not strike; the target remains, but the danger disappears; the danger remains and strikes the agent; and the target disappears, the danger remains but the danger does not strike. When only the target remains, the optimal strategy is to go straight to the target, as in equation (4). When the target disappears but the danger remains, the agent must flee to a safe distance from the danger, as in equation (5). A safe distance is a variable parameter called the danger radius. Once the agent is outside the danger radius, it presumes that it is safe from the danger.

Using the above equations, the expected utility of each grid point in the environment can be calculated provided $EU(O|t, d, \lambda')$ can be accurately determined and θ can be found. Since λ' is most likely between grid-points, the local EU function must be interpolated from the expected utility values of the surrounding grid points. Using the interpolated surfaces for the local values of $EU(O|t, d, \lambda)$, the value of θ can be determined by searching for the angle that maximizes the function described by equation 3. Once the expected utility is determined for a grid point, its value is then used to calculate the expected utility of its neighboring grid-points. This process is repeated until values are collected for all the grid points. Because the estimated utility value can change for a point when the values of its neighbors change, the values of all the points are repeatedly re-estimated until the values stabilize. The pseudocode for the algorithm is given in figure 1.

[1]"Two totally ordered sets $(A, \leq)$ and $(B, \leq)$ are order isomorphic iff there is a bijection from A to B such that for all $a_1, a_2 \in A, a_1 \leq a_2$ iff $f(a_1) \leq f(a_2)$."[Weisstein, 2001]

```
Open is a list of grid-points that need to be updated.
Closed table of updated points.
N is the point currently being updated.
V_N is the current EU estimate at N.
repeat
    Open ← enqueue target locations
    Closed ← ∅
    repeat
        N ← dequeue from Open
        when N ∉ Closed
            EU(O|t, d, λ) ← interpolated EU
                            function from V_N at neighboring points
            V_N ← max equation 3
            Open ← enqueue neighbors of N
            Closed ← add N
    until Open = ∅
until convergence
```

Figure 1: The dynamic programming algorithm for estimating the expected utility at all the grid points in the environment.

3.2 Other Action Selection Mechanisms

It is typically computationally prohibitive for an agent to calculate the optimal action using a technique similar to the one described in the previous section. Instead, many researchers propose easy to compute action selection mechanisms that are intended to approximate the optimal behavior [Cannings and Orive, 1975; Fraenkel and Gunn, 1961; Römer, 1993]. In addition to the optimal strategy described above, we also examine three other action selection strategies:

- Direct: The agent moves directly to the target, ignoring the danger. This is a non-compromise strategy that one would expect to do poorly.
- Max goal: This strategy moves directly to the target unless the agent is within the danger zone. Within the danger zone, the agent moves directly away from the danger until it leaves the danger zone. This strategy zig-zags along the edge of the danger zone as the agent moves toward the target. Max Goal is also a greedy strategy that only acts upon one goal at a time.
- Skirt: This strategy moves directly toward the target unless such a move would enter the danger zone. In this case, the agent moves along the edge of the danger zone until it can resume heading directly to the target. Skirt is also primarily a greedy strategy. Outside the danger radius, the agent moves straight to the target. Inside the danger radius the agent moves straight away from the danger. At the edge of

the danger radius the behavior is still optimal for the avoid danger goal, as any movement not into the danger zone is equally optimal. With respect to the target goal, the movement is sub-optimal.

The expected utility of each of these mechanisms can be calculated for any particular scenario by using equations 3, 4 and 5, where the action θ is the one recommended by the strategy, not the optimal action.

4 Experiments

The experiments were designed to determine how much better the optimal strategy is over the other strategies, as well as qualitatively examine what sorts of compromise actions are exhibited by the optimal strategy. In all of the trials, a target was placed at $(50, 90)$ with a $G_t = 100$ and a danger was placed at $(60, 50)$ with $G_d = -100$. In each trial, a p_t was selected in the range $[0.95; 1)$, and p_d was selected in the range $[0.5; 1)$. The $p_n(d)$ function was one of four functions, all of with with a danger radius of 20:

- Linear A: $p_n(d) = 0.04d + 0.2$ when $d \leq 20$, 1 otherwise.
- Linear B: $p_n(d) = 0.005d + .9$ when $d \leq 20$, 1 otherwise.
- Exponential: $p_n(d) = d^2/400$ when $d \leq 20$, 1 otherwise.
- Sigmoid: $p_n(d) = 1/(1 + 1.8^{10-d})$ everywhere.

Linear A was selected as a baseline strategy where the probability of a strike was high near the danger, but low at the edge of the danger zone. Linear B was selected to make the chance of a strike low overall, thus increasing the tendency to stay in the danger zone longer, for more compromise actions. Exponential has a high probability of a strike for much of the danger zone, but drops off sharply at the edge, perhaps encouraging compromise behavior near the edge. Sigmoid should resemble exponential, but the area with low strike probability is larger, and there is the possibility of some strike for every location in the environment, not just inside the danger radius.

Once the scenario was generated, the expected utility for each of the three non-optimal strategies and the optimal strategy was calculated for 200 points in the environment.

5 Results

Figure 2 shows the results of the optimal strategy when $p_t = 0.995, p_d = 0.99$, and $p_n(d)$ is Linear A. There are two interesting properties to note. First, within the danger zone, there is little display of compromise action; the agent flees directly away from the danger at all locations, ignoring the target. Second, there *is* compromise action displayed outside the danger zone, to the lower right. While this makes sense (the agent would want to take the shortest path around the danger zone) it does not fit into the common conception of compromise action. In that area of the environment, the goal to avoid the danger would not be active (since the agent is too far away from the danger) and thus one would expect it to have no effect of the action selected.

Figure 3 shows the optimal strategy when $p_t = 0.995, p_d = 0.5, p_n(d)$ is Linear A. The main difference is that the compromise action in the lower right is less pronounced. The optimal strategy is to assume that the danger will disappear by the time the agent gets there. This property is seen in all the other experiments, i.e., when p_d is high, the agent avoids the danger zone and exhibits compromise behavior in the lower right region, but when p_d is low, the agent moves straight to the target.

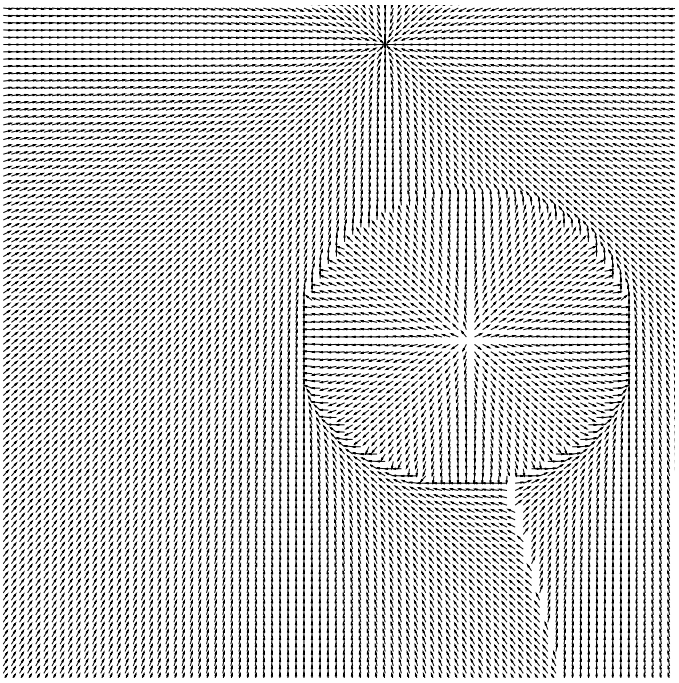


Figure 2: Optimal behavior when the target and danger are likely to stick around ($p_t = 0.995, p_d = 0.99$, and $p_n(d)$ is Linear A).

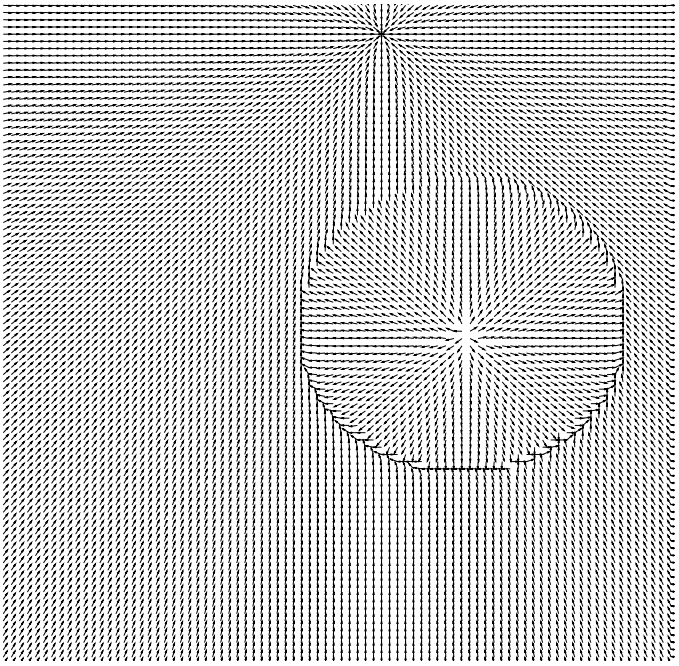


Figure 3: The same scenario as figure 2, but with $p_d = 0.5$. It shows effect of p_d on behavior outside the danger zone.

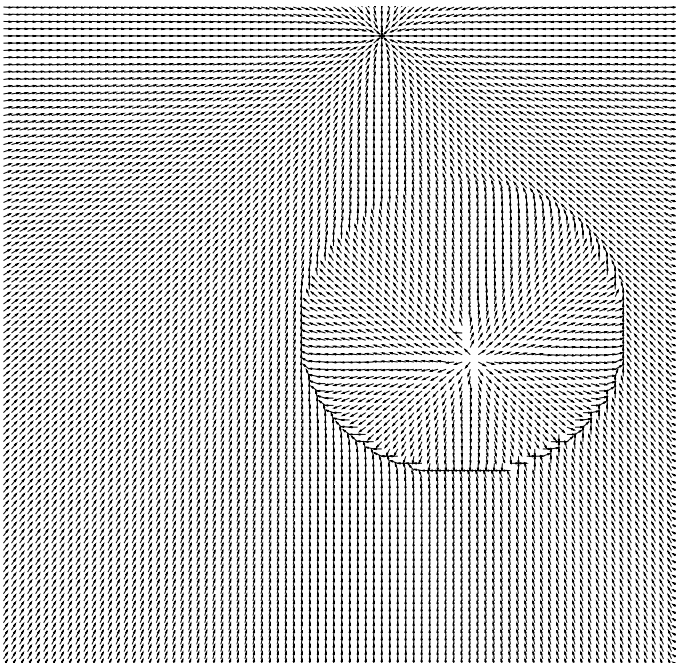


Figure 4: When $p_t = 0.95, p_d = 0.5$, and $p_n(d)$ is Linear A, the results show some compromise action within the danger zone as well as without.

When $p_t = 0.95, p_d = 0.99$, and $p_n(d)$ is Linear A, the results are qualitatively identical to figure 2, but when $p_t = 0.95, p_d = 0.5$, and $p_n(d)$ is Linear A, we start to see some serious compromise action (figure 4). The combination of both the urgency to get to the target with the likelihood that the danger will disappear leads to more target focused behavior in the danger zone.

When using Linear B, the behavior is identical to Linear A when p_d is high. When p_d is low, the low probability of a strike makes the compromise action more pronounced (figure 5).

With the non-linear $p_n(d)$ functions, compromise action is seen clearly in all cases. Figure 6 shows $p_t = 0.995, p_d = 0.99$, and $p_n(d)$ is sigmoid. The compromise behavior is evident both near the center of the danger zone and again near the edges as the probability of a strike drops gradually from the danger. This is the same for the exponential $p_n(d)$.

The quantitative results of the optimal strategy compared to the greedy strategies described above is shown in table 1. The table shows how the various strategies (optimal, max goal, and skirt) compare to each other in term of percentage improvement. The percentages are of the average expected utility for each strategy across all the starting positions and scenarios[2] listed. "All" is across all scenarios and starting positions; "opposite" is across just the starting positions that are opposite from the target (the lower right region); "danger zone" is across the starting positions inside the danger radius; "Linear A" is all positions when the $p_n(d)$ is Linear A; "Linear B"is all positions when the $p_n(d)$ is Linear B; "Exponential" is all positions when the $p_n(d)$ is Exponential"; and "Sigmoid" is all positions when the $p_n(d)$ is Sigmoid". The direct strategy was predictably poor (less than half as good as the other strategies across all trials, and 1/6 as good inside the danger zone) so we omitted those results from the table. We see that across all samples, the optimal behavior performs

[2]A scenario is a single set of values for the parameters in the model.

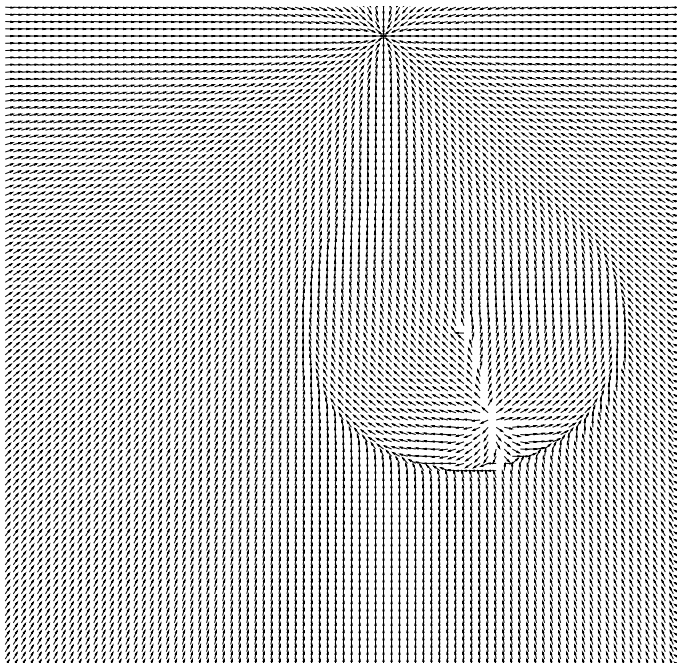


Figure 5: $p_t = 0.95, p_d = 0.5$, and $p_n(d)$ is Linear B.

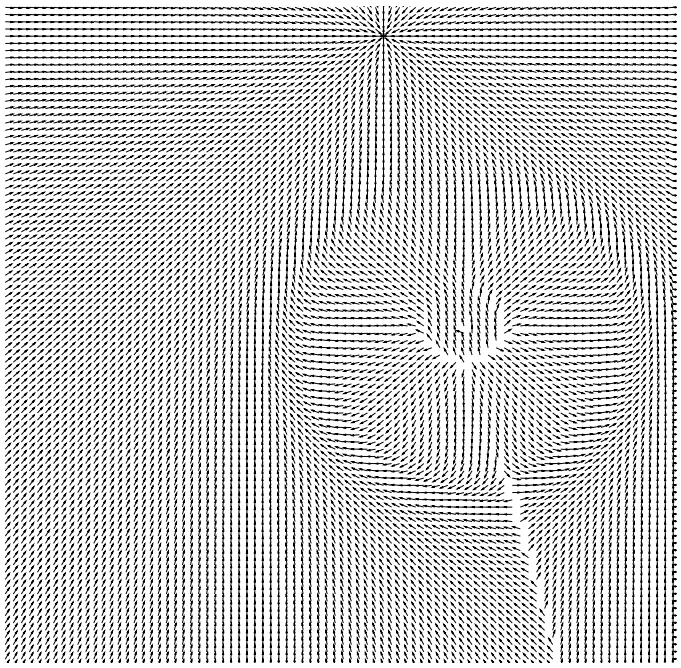


Figure 6: $p_t = 0.995, p_d = 0.99$, and $p_n(d)$ is sigmoid.

scenario	optimal over max goal	optimal over skirt	skirt over max goal
all	29.6%	0.1%	29.1%
opposite	64.9%	0.2%	63.3%
danger zone	26.2%	0.01%	26.1%
Linear A	40.9%	0.02%	40.8%
Linear B	13.5%	0.1%	13.1%
Exponential	48.6%	0.03%	48.5%
Sigmoid	16.7%	0.2%	15.2%

Table 1: Results comparing optimal compromise behavior to the greedy strategies.

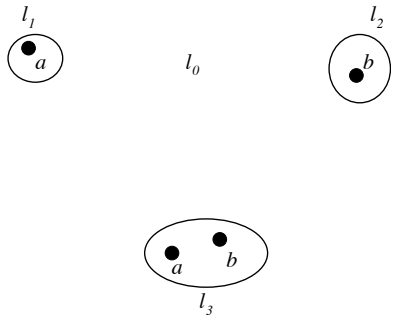


Figure 7: An example scenario where compromise makes sense.

26% better than max goal, but only 0.1% better than skirt. When we consider just those locations on the other side of the danger zone from the target, we see the benefit is greater for optimal over max goal, but still only slightly so over skirt. This is the same for when we consider just those locations inside the danger zone, or we consider just the samples from each of the $p_n(d)$ functions.

6 Discussion

In discussing the results above, we will present some new insights into the nature of the compromise problem, develop its dual nature, propose a new hypothesis and reinterpret the data from ethology.

6.1 Experimental Results

The biggest surprise in the qualitative results is in the number of scenarios where there is almost no compromise action at all. It appears that in stable environments, the priority is to get away from the danger as soon as possible. Even in a case where the target is likely to disappear and the danger unlikely to remain more than a few time steps, with a moderate chance of a strike, the best thing to do is to flee the danger first (figure 4). In contrast, the probability of a strike has a larger effect on the qualitative behavior than we would have suspected, as shown in figures 5 and 6. The pattern of optimal behavior in figure 6 is as we predicted around the edge of the danger zone, but not at all what we expected in the center, with the optimal behavior ignoring the danger entirely. We are exploring possible causes for this.

The quantitative results in table 1 show that compromise actions in the danger zone (an original reason for proposing them) provide much less benefit than compromise actions in the area opposite from the target. On the other hand, the optimal compromise actions are significantly better than the max goal strategy. This arises from the zig-zag nature of this strategy resulting in much longer paths to the target. When this zig-zag is removed (as in the skirt strategy) the optimal strategy is only the slightest bit better. Although there appear to be other patterns in the data with respect to which locations or which $p_n(d)$ functions favor which strategies, more research need to be done to reach a conclusion.

6.2 Blending vs. Voting

In the work so far, we have looked at compromise candidates using the view described in the quotes and examples from ethology given above. The result is that compromise actions qualitatively appear to be blends of the actions best for each sub-goal (the best direction to move is somewhere in-between the directions that are best for each of the goals). There is an alternative description of compromise candidates, also described by Tyrrell, sometimes called the council-of-ministers analogy. In this perspective, there are a collection of "ministers" or experts on achieving each of the agent's various goals. Each minister votes for courses of action that it likes, casting, for example, five votes for its favorite action, four for its second favorite, and so on. The prime minister tallies all the votes and selects the action with the most votes. In this configuration, the compromise selected can be radically different from the non-compromise actions. Imagine an agent at a location l_0 that needs some of resource a and some of resource b. There is a quality source of a at l_1, a location far from a quality source of b at l_2. There is a single low-quality source of both a and b at l_3 (figure 7). Assuming that the utility of a at l_n is a_n, and there is some cost of movement c (a chance of the resource moving away or a direct cost such as energy consumed) then the agent should move l_3 whenever $a_3 + b_3 - c(\overline{l_0, l_3}) > a_1 + b_2 - c(\overline{l_0, l_1} + \overline{l_1, l_2})$. In the council-of-ministers, the a minister would cast some votes for l_1, but also some for l_3. Similarly, the b minister would cast votes votes for both l_2 and l_3. The agent might then select moving l_3 as its compromise choice when it is beneficial.

This presents us with an interesting discrepancy: in one model of compromise selection, drawn from real world examples in ethology, compromise is a form of action blending that appears to have little overall benefit to the agent in the prescriptive goal case, and benefit in a limited sense in the proscriptive goal case. In the other, largely hypothetical, model, compromise seems much more granular, results in actions that are qualitatively different from the non-optimal actions, and appears to have the ability to confer real advantage. In the literature, this contrast (in terms of compromise) is unknown, beginning with Tyrrell who used the two definitions interchangeably.

It is our position that the difference between these two models of compromise are because of the level at which the action is defined. Blending compromises take place at the lower levels, where the outputs are essentially the motor commands for the agent. Thus changes allow for little variation in the output. Voting compromises take place at a higher level, where each choice can result in many varied low-level actions. For purposes of distinction, we will call low-level actions[3] *actions* and higher-level actions[4] *behaviors*[5]. Thus selecting a different behavior module

[3] such as *move 1 unit at 2.1 radians*

[4] such as *go to location* l_3

[5] We rely on the behavior-based robotics notion of "behavior" as a reactive module designed to achieve a particular goal. They are also commonly referred to as *goals* or *tasks*. Their important property is higher level of abstraction over actions.

can have wildly different effects at the action level. We believe this distinction was not made by the early researchers in action selection because their experimental environments were entirely discrete and grid-based, thus affording few action options to the agent. For Tyrrell, there was little difference between compromise actions and compromise behaviors.

We note that the "three-layer architectures" in robotics do explicitly make this distinction, where higher layers select between multiple possible behaviors, and then at lower layers, multiple active behaviors select actions [Gat, 1991; Bonasso *et al*., 1997]. When and where compromise behavior is included varies from instance to instance in an ad hoc manner. Many modern hierarchical action selection mechanisms that explicitly use voting-base compromise tend to do so at the behavior level only [Pirjanian *et al*., 1998; Pirjanian, 2000; Bryson, 2000].

6.3 The Compromise Behavior Hypothesis

The experiments here and in previous work, with the insights discussed above, lead us to propose the following Compromise Behavior Hypothesis:

> Compromise at the action level confers less overall benefit to an agent than does compromise at the behavior level. Compromise behavior is progressively more useful as one moves upward in the level of abstraction at which the decision is made, for the following reasons:
>
> 1. In simple environments (e.g. two prescriptive goals), optimal compromise actions are similar to the possible non-optimal compromise actions as well as the possible non-compromise actions. As such, they offer limited benefit. In these environments there is no possibility of compromise at the behavior level.
> 2. In complex environments (e.g. where multiple resources are to be consumed in succession such as the scenario depicted in figure 7) compromise behavior can be very different from the active non-compromise behaviors, endowing it with the potential to be greatly superior to the non-compromise.
> 3. In complex environments, optimal or even very good non-optimal actions are prohibitively difficult to calculate.

In the complex environments, optimal compromise *actions* may offer little benefit over actions derived from compromise *behaviors* for the same reason as in 1 above: the optimal action is too similar to the non-optimal action. For example, in the figure 7 scenario, a behavior that decides to move the agent to l_3 can ignore the locations of a and b at l_3 and just generate an action to move to l_3 in general. This non-optimal behavior-generated action will be nearly as good as the optimal action generated by considering the location and qualities of all the a and b, yet the optimal action will come at an enormous computational cost. We propose to begin testing this hypothesis with just this scenario. We predict that the optimal action will be to move toward a location between a and b in l_3, but this optimal action will be essentially just as good as a movement to any other part of l_3.

6.4 Ethological Data Reinterpreted

If it is true that compromise actions are less helpful than compromise behavior, why are so many examples drawn from ethology used to demonstrate compromise actions in animals? It may be that the interpretation of the animal data has been overzealous. In each case there are other possibilities to explain the behavior that do not involve the weighing of compromise actions, or even involve the animal's action selection mechanism at all.

For instance, in the two-prescriptive-goal examples with frogs and crickets following a curved path between two prescriptive goals, an alternative explanation might be that the multiple targets are being merged at the perceptual level, with the ear or auditory system averaging the position of the two targets before any action selection mechanism has an opportunity to consider its options. In this interpretation the behavior would be an accident of morphology, not an attempt to maximize the creature's utility.

Some examples of potential compromise behavior, such as dogs combining a display of fear with one of anger [Lorenz, 1981], or the goose trembling when torn between a prescriptive and a proscriptive goal, may be less an example of compromise behavior, and more a superposition of the two behaviors. This effect arises from the behaviors not sharing a common final path, enabling them to both be expressed simultaneously. In the case of the geese, the resulting behavior is probably one of the least beneficial actions that could be selected, rather than approximating optimality.

Admittedly, these re-interpretations are speculative, but they may not be much more speculative than the idea that they are a result of deliberate consideration of compromise candidates.

7 Conclusions and Future Work

In this paper we have analyzed the properties of action selection mechanisms in a scenario that has been of interest to both ethologists and AI researchers in the past. In it we have shown that optimal compromise actions in a proscriptive goal case are qualitatively different from what was predicted. They further afford little benefit when compared to a minimally compromise-enabled strategy. We proposed that compromise is not especially useful at the action level, but is useful at the higher behavior level. Future work will revolve around testing, validation or refutation of this Compromise Behavior Hypothesis.

Acknowledgements

We would like to thank Chris Brown and Rebecca Hwa for many wonderful discussions. We would also like to thank the anonymous reviewers for several helpful comments. This work was sponsored in part by a grant from the Office of Naval Research, number N0001404WR20377.

References

[Arkin, 1998] R. Arkin. *Behavior-Based Robotics*. MIT Press, Cambridge, MA, 1998.

[Avila-Garcia and Canamero, 2004] O. Avila-Garcia and L. Canamero. Using hormonal feedback to modulate action selection in a competitive scenario. In *From Animals to Animats 8: Proceedings of the Eighth International Conference on Simulation of Adaptive Behavior (SAB)*, pages 243–254, 2004.

[Bailey *et al*., 1990] W. J. Bailey, R. J. Cunningham, and L. Lebel. Song power, spectral distribution and female phonotaxis in the bushcricket requena verticalis (tettigoniiddae: Orthoptera): active femail choice or passive attraction. *Animal Behavior*, 1990.

[Blumberg, 1994] B. M. Blumberg. Action-selection in hamsterdam: Lessons from ethology. In *Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, 1994.

[Bonasso *et al*., 1997] R.P. Bonasso, R.J. Firby, E. Gat, D. Kortenkamp, and D. Miller. Experiences with an architecture for intelligent, reactive agents. *Journal of Experimental and Theoretical Artificial Intelligence*, 9(2):237–256, 1997.

[Brooks, 1986] R. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, RA-2:14–23, 1986.

[Bryson, 2000] J. Bryson. Hierarchy and sequence vs. full parallelism in action selection. In *Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*, 2000.

[Caldwell, 1986] G. Caldwell. Predation as a selective force on foraging herons: effects of plumage color and flocking. *Auk*, 103:494–505, 1986.

[Cannings and Orive, 1975] C. Cannings and L. M. Cruz Orive. On the adjustment of the sex ratio and the gregarious behavior of animal populations. *Journal of Theoretical Biology*, 55:115–136, 1975.

[Choset *et al*., 2005] H. Choset, K. M. Lynch, S. Hutchinson, G. Kantor, W. Burgard, L. E. Kavraki, and S. Thrun. *Principles of Robot Motion: Theory, Algorithms, and Implementations*. MIT Press, Cambridge, MA, 2005.

[Crabbe and Dyer, 1999] F. L. Crabbe and M. G. Dyer. Second-order networks for wall-building agents. In *Proceedings of the International Joint Conference on Neural Networks*, 1999.

[Crabbe, 2004] F.L. Crabbe. Optimal and non-optimal compromise strategies in action selection. In *Proceedings of the Eighth International Conference on Simulation of Adaptive Behavior*, 2004.

[Decugis and Ferber, 1998] V. Decugis and J. Ferber. An extension of maes' action selection mechanism for animats. In *Proceedings of the Fifth International Conference on Simulation of Adaptive Behavior*, 1998.

[Fraenkel and Gunn, 1961] G. S. Fraenkel and D. L. Gunn. *The Orientation of Animals*. Dover, New York, 1961.

[Fraser and Cerri, 1982] D.F. Fraser and R.D. Cerri. Experimental evaluation of predator-prey relationships in a patchy environment: consequences for habitat use patterns in minnows. *Ecology*, 63, 1982.

[Gat, 1991] E. Gat. *Reliable Goal-directed Reactive Control for Real-world Autonomous Mobile Robots*. PhD thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1991.

[Girard *et al*., 2002] B. Girard, V. Cuzin, A. Guillot, K. N. Gurney, and T. J. Prescott. Comparing a brain-inspired robot action selection mechanism with 'winner-takes-all'. In *Proceedings of the seventh international conference on simulation of adaptive behavior: From animals to animats*, pages 75–84, 2002.

[Grubb and Greenwald, 1982] T.C. Grubb and L. Greenwald. Sparrows and a brushpile: foraging responses to different combinations of predation risk and energy cost. *Animal Behavior*, 30, 1982.

[Howard, 1977] R.A. Howard. Risk preference. In *Readings in Decision Analysis*. SRI International, Menlo Park, CA, 1977.

[Humphrys, 1996] M Humphrys. *Action Selection methods using Reinforcement Learning*. PhD thesis, University of Cambridge, 1996.

[Hutchinson, 1999] J.M.C. Hutchinson. Bet-hedging when targets may disappear: optimal mate-seeking or prey-catching trajectories and the stability of leks and herds. *Journal of Theoretical Biology*, 196:33–49, 1999.

[Jones *et al*., 1999] R. M. Jones, J. E. Laird, P. E. Nielsen, K. J. Coulter, P. Kenny, and F. V. Koss. Automated intelligent pilots for combat flight simulation. *AI Magazine*, 20(1), 1999.

[Latimer and Sippel, 1987] W. Latimer and M. Sippel. Acoustic cues for female choice and male competition in the tettigonia cantans. *Animal Behavior*, 1987.

[Lorenz, 1981] K. Z. Lorenz. *The Foundations of Ethology*. Springer-Verlag, New York, 1981.

[Maes, 1990] P. Maes. How to do the right thing. *Connection Science Journal, Special Issue on Hybrid Systems*, 1, 1990.

[Milinksi, 1986] M. Milinksi. Constraints places by predators on feeding behavior. In T.J. Pitcher, editor, *The behavior of teleost fishes*, pages 236–256. Croom Helm, London, 1986.

[Montes-Gonzales *et al*., 2000] F. Montes-Gonzales, T. J. Prescott, K. Gurney, M. Humphrys, and P Redgrave. An embodied model of action selection mechanisms in the vertebrate brain. In *Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*, pages 157–166, 2000.

[Morris *et al*., 1978] G. K. Morris, G. E. Kerr, and J. H. Fullard. Phonotactic preferences of female meadow katydids. *Canadian Journal of Zoololgy*, 1978.

[Pirjanian *et al*., 1998] P Pirjanian, H.I. Christensen, and J.A. Fayman. Application of voting to fusion of purposive modules: An experimental investigation. *Journal of Robotics and Autonomous Systems*, 1998.

[Pirjanian, 2000] P. Pirjanian. Multiple objective behavior-based control. *Journal of Robotics and Autonomous Systems*, 2000.

[Römer, 1993] H. Römer. Environmental and biological constraints for the evolution of long-range signalling and hearing in acoustic insects. *Philosophical Transactions of the Royal Society B*, 340:179–185, 1993.

[Tyrrell, 1993] T. Tyrrell. *Computational Mechanism for Action Selection*. PhD thesis, University of Edinburgh, 1993.

[Weisstein, 2001] E. W. Weisstein. *Eric Weisstein's World of Mathematics*. http://mathworld.wolfram.com/, 2001.

[Werner, 1994] G. M. Werner. Using second order neural connection for motivation of behavioral choices. In *Proceedings of the Third International Conference on Simulation of Adaptive Behavior*, pages 154–161, 1994.

He/She-You-I Formalism: A Heuristic Model of (*En*)Action to Make Decisions

Daniel Mellet-d'Huart
AFPA DEAT / Université du Maine LIUM
dmdh@dm-dh.com

Abstract

This paper presents a heuristic based on a *model of (en)action*. The underlying model of action is rooted in research in neurosciences and based on a conceptual framework. It is organized as a three-pole system within the framework of the "He/She-You-I formalism." The "He/She Pole" deals with understanding, simulation and anticipation of action; the "You Pole" with decision-making and taking engagement within action; and the "I Pole" with executing action within a particular environment.

Key-words

Action, enaction, model, design method, heuristic

1. Introduction and context

This paper presents a *heuristic model of (en)action* [Mellet-d'Huart, 2004] that stands for *embedded* and *embodied action*. "*Action*" encompasses both outer and inner processes a human being engages when acting. Therefore, a particular formalism is proposed in order to support action modeling. Our model of (*en*)action is a global and systemic model. Theory supporting this model comes mainly from Berthoz [1997, 2003], Maturana and Varela [1980]. Because the model links action with body activity in a coherent way with neurophysiology of action and biology of cognition, we refer this model of action to the theory of "*enaction*" [Varela *et al.*, 1991; Varela, 1994]. Thereby, we call it *model of (en)action.*

This model of (*en*)action was formerly developed in order to support and facilitate the design of virtual environments for learning. In the field of virtual reality, Durlach and Mavor [1995] encouraged the setting up of conceptual models based on cognitive sciences to support the design of application dealing with games, education and training. Winn [2003] proposed that learning in artificial environments could be enhanced if more attention was given to embodiment, embeddedness and dynamic adaptation [Winn, 2002]. Being developed as a heuristic, it might have different applications and support action modeling for software applications. In this context, we present this model in this paper.

2. A model of (*en*)action

The model of (*en*)action is supported by a conceptual framework. The conceptual framework integrates generic operators. It is based on: (1) a dynamic *cycle of action*; (2) a "*He/She-You-I" Structure*; (3) a set of three basic-processing operators, *actualization*, *potentialization*, and *virtualization*. This framework constitutes a *formal system* that can be used dynamically, iteratively and be organized in different levels. The "He/She-You-I formalism" is intended to support a *model of action.*

2.1. A dynamic action cycle

The *He/She-You-I Structure* is supported by neurophysiological theories. For Berthoz [1997, 2003], action can be discomposed in three different steps: (a) Anticipation of possible action and its foreseen consequences on the environment; (b) Decision-making and engagement of action; (c) Execution of the chosen action in the extern environment of the human being ("real world").

(a) Before acting in a particular context, a human being explores possible actions by simulation. He/she anticipates expected consequences of action. Therefore, he/she uses models that he/she can apply on new situations. Those models result from formers actions. Data is abstracted from previous experience and conceptualized. Memory registers those data and provides conceptualized elements in order to support anticipations and to produce predictions for new situations. This process minimizes contextual elements and enhances structural or unchanging elements.

(b) A human being has to preserve his/her existence as a living being. Therefore, he/she controls and evaluates both his/her internal and emotional states, and risks, which exist in his/her environment. Previous experiences constitute his/her frame of reference. Different levels of inner states can be distinguished, ranged from visceral internal state to high-level mental states. This stage of action is characterized by evaluating the current risks by scanning the environment and searching for significant details. Thereby, decision-making occurs to be mostly irrational and an

accurate anticipation can be rejected because of an inner feeling of being insecure.

(c) Executing an action involves to engage body-movements. It often aims on inducing intentional changes in the environment, or on protecting and maintaining safe his/her body. Moving body means to deal with inertia, gravity, equilibrium, evaluating distances and forces to be engaged. This is based on imitation, experience and practice. Therefore, one has to develop internal markers and external markers (mainly visual cues).

2.2. A "He/She-You-I formalism"

All components and aspects of the human life are linked together and rooted in biological foundations. Thereby, we join Maturana and Varela [1980] when they approach biology of cognition through embodiment. Following their claims, we consider that language has biological roots [Maturana *et al.*, 1995] and encloses basic universal structures. Our heuristic model is organized following the grammatical distinction of three pronouns: "I", "You", and "He/She." The *He/She-You-I formalism* supports a threefold approach of the human being. It articulates action, embodiment, mind, environment, and contextual factors. The underlying assumption is that human activity can be more easily modelized if three main functions are distinguished. Those functions are internal and co-exist whichever activity is engaged.

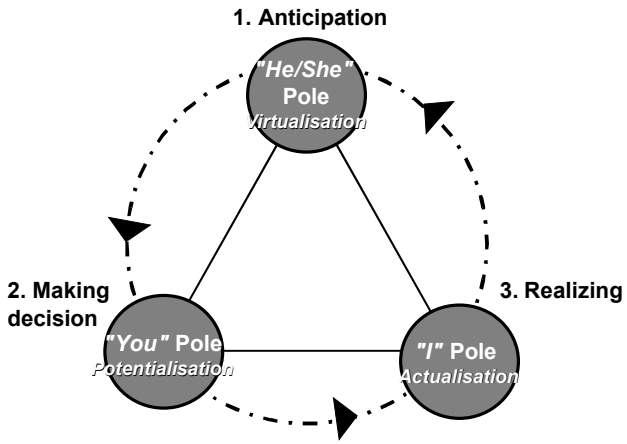


Figure 1 : Conceptual framework: The "*He/She-You-I" formalism*

(a) The *"He/She Pole"* supports anticipation and simulation of action and its expected consequences. It deals with abstraction, causation, generalization, and conceptualization. It provides means for observing other persons' actions in order to better understand how to act on the world. As neuro-physiologists explain, he/she may unconsciously or consciously imitate other persons. It refers to cognition, from acquiring causation to abstract reasoning. Thereby, its main process is "*virtualisation*." The "He/She pole" time if hierarchical and chronological in order to support causation. Events are organized from past to future. But one is free to move and to explore past, present and future. Past experiences are used to simulated action and anticipate possible consequences of action. Space is approached as a Euclidean space; it is flat and can be abstractly manipulated. We identify its main process as "*virtualization*".

(b) The "*You" Pole* deals with decision-making and is interested in maintaining the organism equilibrium [Maturana, 1980; Maturana *et al.*, 1995]. Therefore, it will take into consideration emotional internal states, and do not refer to rational considerations. It is deeply rooted in the internal existence of the organism. It focuses on internal states, survival issues, welfare and relationships with pairs. It supports the engagement that will be required to execute the action. That is where *emotioning* [Maturana 1980] occurs based on inner states. In a certain way, the "You pole" is timeless. It is anchored in natural and physiological cycles. What once was a threat may remind one of the threats forever. It has nothing to do with historical time, but rather with physiological cycles and symbolic functions of events regarding one's own life. It starts, develops and ends things; then restarts, re-develop and re-ends things, and so on. Space is curve; bringing back to starting point. The "You pole" is meshing emotions with space-time events. In this model, the process engaged by decision-making is identified as "*potentialization.*"

(c) The "*I Pole*" focuses on executing adequately targeted and forecasted actions in order to produce expected changes or effects in the environments. The "*I Pole"* commands acting in the external environment of the organism. It organizes body-movement within space. Therefore it is based on the development of sensory-motor skills. By observing and imitating, he/she becomes an "I" who can act by he/herself, develops actions and makes realizations. This three linguistic forms of subjects, underlies the definition of three different poles on which we found (*en*)action. The "I pole" is about *actualizing*. It is an actor position. Thanks to this pole, action may become actual through the mobilization sensitive and effectrice surfaces required by executing. The organism develops movements to achieve its goals. The "I pole" time is rooted in present. What occurs, occurs *now*, within a temporal window for possible actions on the world. Space is condensed and imitated to *here*, within the spatial window for possible actions on the world. Space-time is contracted on here and now in a particular context. Its dynamics is strength. The generic operator of realization is "*actualization*".

The "He/She-You-I formalism" consists in three different poles based on three different processes. Those three different poles co-exist and work together to produce embedded and embodied action. That is what we call (*en*)action. Each pole is correlated with one of the three steps of the neurophysiological action model.

2.3. Action coordination

A vertical organization of our model defines the *phenomenal space of living* [Merleau-Ponty, 1971] for an organism on his/her environment. It takes its place within our polar structure (horizontal component of the structure).

Vertically, it takes place from: (1) physiological anchorage, embodiment and biological grounding (downward), (2) to wholeness a consciousness (upward) [Damasio, 2001, 2002; Delacour, 2001; Mazoyer, 2002]. It links together perception, action and phenomenological experience [Noe and O'Regan, 2000; Myin and O'Regan, 2002]. Within this framework different kind of action may take place from low-level movements to highly complex and long-term activities. Intermediate levels facilitate the consideration of coordination activities and growing part taken by conceptual and formal activities. Thereby, although it remains as a difficult and quite formalized aspect of the enactive model of action, it opens perspective to bridge and create passageway between different theories as [Piaget, 1974a, 1974b, Maturana *et al.*, 1995; Damasio, 2001, 2002]. It has to be pointed out that those different theories establish vertical and progressive organization in different levels to explain different forms of engagement in simple body activity Vs conceptual and complex language activity. Thereby, this model provides bases to support research on mechanisms enabling to pass from one level to another (coordination, becoming aware of, conceptualization…). It facilitates the breaking down to parts and the reverse movement to make up the whole again. It constitutes a vertical and dynamic structure. An organization within different levels of action coordinating (going upward, more and more complex actions occur) is used to support choices about coupling users and virtual environments in design situations. Within this phenomenal space of living, we can refine distinctions up to seven levels.

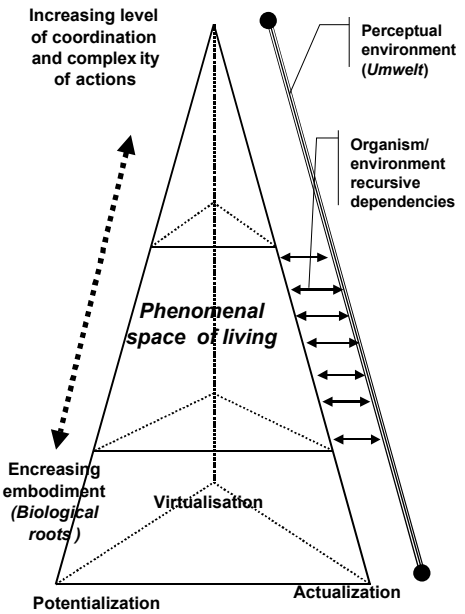


Figure 2 – Vertical structure and organism/environment coupling

3. Choosing what to do next

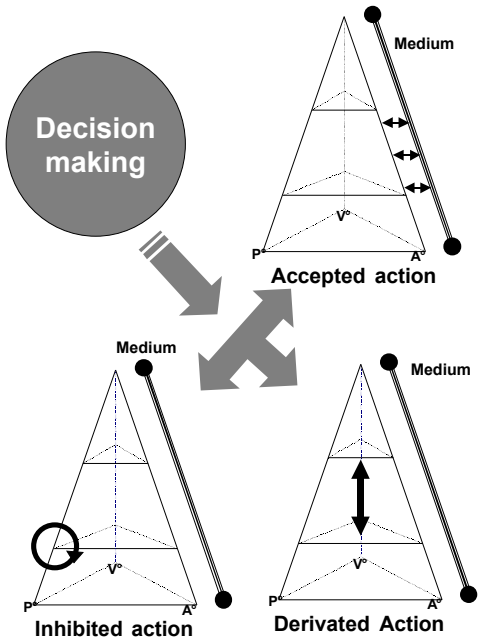


Figure 3 – Decision-making about acting: three possible outcomes

In order to choose what to do next, both the "He/she pole" and the "You pole" are important. When the former provides anticipation of what can be done next, the "You pole" will make decision of which anticipation has to be engaged and/or inhibited. The purpose of decision-making is to select the actualized outcome of action. Decision-making is the purpose of the "You pole." Thereby, it treats the simulations of possible consequences of action that are produced by the "He/She pole." Traditional approaches of decision present two types of outcome of decision: a simulated action is *accepted* or *inhibited*. Current neurosciences approaches underline the role of inhibition [Berthoz, 1997, 2003; Houde *et al.* 2002]. Pre-motor theories as well as mirror-neuron approaches [Kohler *et al.*, 2002; Rizzolatti and Craighero, 2004; Gallese, 2005] enlighten how neurophysiological phenomena are engaged whether actualization of action is decided or not. The explanation is that often active inhibition is required not to act. *Acceptation of action* will engage the execution or *actualization* of action ("I pole") based on anticipations produced by the "He/She pole." This mobilizes energy and has to be spatially adapted. Differently, *inhibition of action* leads to no changes in the outer world. Actualization of action might be either cancelled or postponed. Another temporal localization may have to be found and energy may have to be recovered or accumulated. As far as *diversion of action* is concerned, an enlargement of mental views or mechanisms has to be found. Thereby, we distinguish three possible outcomes for decision-making as shown in the previous figure.

4. Discussion and perspectives

We presented th e outline of a heuristic Model of (*En*)Action and its *He/She-You-I forma lism* [Mellet-d'Huart, 2004]. This model may help to renew the *mod el of decision* of *autonomou s agents*. It suppo rts an embodied and embedded approach of acting and learning. Even if the former conceptual model of autonomous agent activity (perception-decision-action cycle) might seem to be cl ose and can also be cut in three steps, the model of (*En*)Action differs radically from the p revious. In one case, it is a computational model (*perception* = inp ut of information; *decision* = op erating a treatment on data; *action* = output on environment). In an enactive model, *anticipation* is a complex and active process based on body experience and former conceptualization; *decision m aking* consists in evaluating internal p hysiological and em otional state Vs emotional connotation of external data; and *executing* deals with a complex engagemen t of a finalized b ody activity that aims to a change in a c ontextual environment. Each of those three steps involves a coor dination of perceptive and motor skills. O nly the focus of this ac tivity changes depending of where we are within the action cycle.

References

[Berthoz, 1997] Berthoz, A. *Le sens du mouvement.* Odile Jacob, 199 7.

[Berthoz, 2003] Berthoz, A. *La décision.* Odile Jacob, 2003 .

[Damasio, 2001] Damasio, A. R. *L'erreur d e Descartes.* Odile Jacob, 200 1.

[Damasio, 2002] Damasio, A. R. Le sentiment d'être soi. Corps, émotions, conscienc e. Odile Jacob, 200 2.

[Delacour, 2001] Delacour, J. *Conscience et cerveau : la nouvelle frontière des neurosciences.* DeBoeck Université, 200 1.

[Durlach and Mavor, 1995] Durlach, N. and Mavor, A. Eds. *Virtual reality. Scientific and technolog ical challenges.* Naval training systems center 1995 .

[Gallese, 200 5] Gallese, V. Embodied simulation: From neurons to phenomenal experience. Phenomenology and the cognitive science. 2005. http://www .unipr.it/arpa/mirror /english/staff/gallese.htm.

[Houde *et al.*, 2002] Ho ude, O., Mazoyer, B. and Tzourio-Mazoyer, N. Eds. *Cerveau et psych olog ie. Introduction imag erie cérébrale anatom ique et fonctionnelle.* PUF, 2002.

[Kohler *et al.*, 2002] Kohler, E., Keysers, C., Umilta, M.A., Fogassi, L., Gallese, V. and V.Rizzolatti, G. Hearing sound, un derstanding actions: action representation i n mirror neurons. Scie nce. 297 , 846-848. 2002.

[Maturana and Varela, 198 0] Maturana, H. and Varela, F. *Autopoiesis and Cognit ion: The Realization of the Living,* Reidel Publish ing, 198 0.

[Maturana, 198 0] Maturana, H. R. Biology of c ognition. In Dordecht, D. Ed. *Autopo iesis and Cogn ition: The Realization of th e Living*. pp. 5-58, Reidel Publishing, 1980.

[Maturana *et al.*, 1995] Maturana, H. R., Mpodozis, J. and Letelier, J. C. Br ain language and the origin o f human mental functions Biological research 28 pp .15-26 www.informatik.u mu.se/~rwhit/MatMpo&Let(1995).ht ml, 1995.

[Mazoyer, 200 2] Mazoyer, B. La c onscience. In Houde, O., Mazo yer, B. and Tzourio-Mazoyer, N. Eds. *Cerveau et psycholog ie. Introduction imag erie cérébrale ana tomique et fonctionnelle*. PUF, 2002.

[Mellet-d'Huart, 2004] Mellet-d'Huart, D. *De l'intention à l'attention : contributions à une démarche d e conception d'environ nements virtuels pou r apprendre à pa rtir d'un mo dèle de l'(én)action*. PhD. T hesis. Université du Maine. 2004. http://www.dm-dh.com/Publi&Achievments_EN.htm

[Merleau-Ponty, 1971] Merleau-Ponty. *Existence et dialectique.* Presses Universitaires de France, 19 71.

[Myin and O'Regan, 200 2] Myin, E. and O'Regan, K. Perceptual consciousness access to modality and skill theories: A way to naturalise pheno menology? Journ al of consciousness studies 9 No. 1 20 02 pp. 27-45, 2002.

[Noe and O'Regan, 2000] N oe, A. and O'Re gan, K. Perception attention and the grand illusion. In Attentional Blindness. Psyche 6(15) October 200 0 http://ps yche.cs.monash.edu.au/v6 /psyche-6-15-noe.html, Arien Mac k's and Irvin Rock's Book, 200 0.

[Piaget, 1974 a] Piaget, J. La prise de conscience. PUF, 1974.

[Piaget, 1974b] Piaget, J. Réussir et comprendre. PUF, 1974.

[Rizzolatti and Craighero, 2004] Rizzolatti, G. and Craighero, L. The mirro r neuron system. Ann. Rev. Neuroscience. 27; 169-192. 2004.

[Varela *et al.*, 199 1] V arela, F. J., Thompson, E. a nd Rosch, E. The Emb odied Mind: Cognitive Science and Huma n Exp erience. 1991.

[Varela, 19 94] Varela, F. J. Autopoiesis and a biology of intentionality In Mc Mullin and Murph y Eds. Autopoiesis and Perception. Schoo l of Electronic Engineering, 199 4.

[Winn, 2002] Winn, W. Learning in Artificial Environments: Embo diment Embeddedness an d Dynamic Adaptation. In Techniques Instruction Cognition and Learning. Vol. 1 20 02. Old City Pub lishing, Inc., 2002.

[Winn, 200 3] Winn, W. Beyond constructivis m: A return to Science-based research and p ractice in Educational Technology, 200 3.

Forced moves or good tricks in design space?
Great moments in the evolution of the neural substrate for action selection

Tony J. Prescott
Adaptive Behaviour Research Group,
University of Sheffield, UK.
t.j.prescott@sheffield.ac.uk

Abstract

This mini-review considers some important landmarks in the evolution of animals, asking to what extent specialised action selection mechanisms play a role in the functional architecture of different nervous system plans, and looking for 'forced moves' or 'good tricks' (Dennett, 1995) that could possibly transfer to the design of control systems for mobile robots. A key conclusion is that while *cnidarians* (e.g. jellyfish) appear to have discovered some good tricks for the design of behaviour-based control systems—lacking specialised selection mechanisms; the evolution of bilaterality in *platyhelminthes* (flatworms) may have forced the evolution of a central ganglion (or 'archaic brain') whose main function is to resolve conflicts between competing peripheral systems. Whilst vertebrate nervous systems contain many interesting substrates for selection it is likely that here too, the evolution of centralised selection structures such as the *basal ganglia* and *medial reticular formation* may have been a forced move due to the need to limit connection costs as brains increased in size.

1 Introduction

Action selection is the task of resolving conflicts between competing behavioral alternatives. This problem has received considerable attention in the growing adaptive behavior literature (see Maes, 1995; Prescott, Redgrave, & Gurney, 1999) much of which has built on earlier research in ethology (see e.g. McFarland, 1989) where it is also described as the task of 'decision making', 'behavior selection', or 'behavior switching'. Whichever label is used, it is useful to recognise at the outset that the problem of selecting actions is really part of a wider problem faced by any complete creature, that of behavioural integration—

"the phenomenon so very characteristic of living organisms, and so very difficult to analyse: the fact that they behave as wholes rather than as the sum of their constituent parts. Their behaviour shows integration, […] a process unifying the actions of an organism into patterns that involve the whole individual." (Barrington, 1967, p. 415)

In discussing control systems for mobile robots, Brooks (1994) has emphasised a similar notion of behavioral coherence which he places at the centre of the problem of autonomous agent design. As robots have become more complex, they have naturally gained an increasing variety of actuator sub-systems, many of which can act in parallel. Controlling robots therefore requires the co-ordination, in space and time, of many interacting sub-systems, and the allocation of appropriate resources between them. The problem for control system design is to satisfy these multiple constraints in a manner that maintains the global coherence of the robot's behaviour. Given this context, Brooks raises the concern that research directed at the more specific problem of action selection may not lead to automatic progress in the design of systems with behavioral coherence. It may be the case, for instance, that proposed action selection mechanisms will not scale-up to the task of controlling more complex robots; or, that we may come to see effective action selection as a consequence of maintaining behavioral coherence, rather than as a key element involved in creating it.

What concerns us here, of course, is the question of the decomposition of control, also known as the 'problem of architecture'. Will an effective robot controller have components whose role is recognisably to resolve conflicts between different action sub-systems? Or, is action selection better regarded as an emergent property—the consequence of many and diverse interactions between multiple sub-systems? (and, in this sense, not something to be considered in isolation from other aspects of control). If effective integration is emergent then research on the design of action selection mechanisms *per se* may lead to a dead-end. On the other hand, if action selection or other related aspects of behavioural integration, can be implemented in specialised system components, then some of the advantages of modularity may accrue to the whole design process. Specifically, it may be possible to add/delete/modify different action sub-systems with less concern for the possibility of adverse, system-wide consequences for the maintenance of behavioural coherence.

How are we to decide answers to these questions? Our strategy in this paper is to attempt a brief survey of some relevant characteristics in the design of natural control systems for complete creatures—animal nervous systems. Our focus will be on those aspects of the functional architecture of nervous systems that seem to play an important role in action selection, or, more

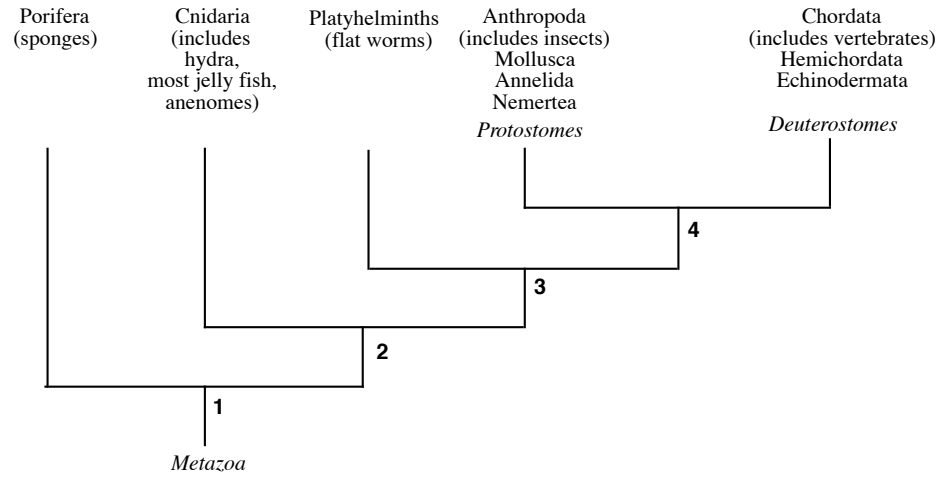


Figure 1. Phylogeny of early metazoans (based on Raff, 1996). Important evolutionary changes include 1. multicellularity, 2. Radial symmetry, different tissue types, nerve nets. 3. Bilateral symmetry, internal organs, central nervous systems, brains. 4. *Hox* gene expression in nervous system and body patterning.

broadly, in behavioural integration. In particular, we will look for evidence of structures that are specialised to resolve conflicts and that seem to have this as their primary function. The absence of such structures would favour the view that action selection is most often the emergent consequence of the interaction of sub-system elements concerned with wider or different aspects of control. Such findings might encourage us to pursue similar, distributed solutions to the coordination of complex robot control systems. The presence of candidate structures, on the other hand, would favour the view that complex control architectures can have a natural decomposition into components concerned with the sensorimotor control of action, and those concerned with the selection of action. Such findings would suggest a similar strategy for the decomposition of robot control. To anticipate our argument somewhat, we will be making the case that nervous system evolution does show evidence of specialised action selection mechanisms in some complex natural control systems.

Our approach is also an evolutionary one in that we will specifically consider animal nervous systems at three different and important grades in the evolution of complex metazoans (multicellular animals). To place what follows in this evolutionary context, figure 1 shows a phylogeny of the major metazoan phyla illustrating some of the principle early events in the evolution of animal body plans and nervous systems.

From the perspective of this paper, the first event of particular note is the evolution (node 2 in the figure) of neurons and nerve nets in animals of the phylum *cnidaria*. This phylum includes a host of relatively simple, but also very intriguing animals such as jellyfish, sea anemones, corals, and hydrozoa (e.g. Hydra). These differ from the most primitive metazoa (the sponges—*Porifera*), in that they possess a variety of different tissue types; generally possess a radial symmetry; may have simple sensory organs; and have nervous systems composed of networks of nerve cells. Fossil evidence suggests that cnidaria were present in the Precambrian era (i.e. more than 550 million years ago), and are therefore likely to have been the first animals to evolve nervous systems of any kind. There is still a great deal to be learned about the functional architecture of cnidarian nervous systems, however, existing research does provide a number of very interesting pointers. Some of this evidence is reviewed in section 2 below.

The next event, node 3, separates the bilateral animals from the other metazoan phyla, and identifies the *platyhelminthes* (the flatworms) as the most primitive form of bilaterian. Flatworms possess central nervous systems organised around a 'brain'. Animals of this sort are known to have been present in the Precambrian as demonstrated by the large number of trace fossils that have preserved the behaviour (e.g. foraging trails) though not the body forms of worm-like animals from that period. Simulation of these trace fossil patterns indicates a capacity for intelligent coordinated behaviours not unlike that demonstrated in some simple behaviour-based robots (Raup & Seilacher, 1969; Prescott & Ibbotson, 1997). Section 3 reviews a number of findings concerned the functional architecture of the nervous systems of living platyhelminthes.

Node 4 in our figure marks the beginning of a further momentous phase—the evolution of the metazoan phyla who share the use of the Hox regulatory gene cluster as a determinant of body patterning and nervous system organisation. Many diverse animal types are listed here that can be distinguished into two distinct groups, the *protostomes* and the *deuterostomes*, on the basis of early events in embryological development. It is interesting to note that the evolutionary line leading to the vertebrates (belonging to the phylum *chordata*), probably diverged at a very early stage from that leading to invertebrate groups with more 'advanced' nervous systems (insects, cephalopods, etc.)—the common ancestor of all these bilaterians being of only flatworm grade. Of the deuterostomes, in fact, vertebrates are the only animals with highly developed nervous systems, although the *echinoderms*—such as sea urchins, and starfish—with their pentameral (five-sided) symmetry present some interesting problems (and solutions) in control system design!

Fossil evidence shows a remarkable explosion of animal forms during the Cambrian period (543–505 million years ago) in which all of the more advanced protostome and deuterostome phyla were represented, having been almost entirely absent from the fossil record at the end of the Precambrian. This evidence suggests the very rapid evolution of complex nervous systems as part of the general evolution of new body plans (Gabor Miklos, Campbell, & Kankel, 1994). Until relatively recently there were no uncontroversial vertebrate fossils of earlier origin than the Ordovician period (~495 million years ago), implying that vertebrates appeared somewhat later than this general explosion of bilaterians. However, finds from Chengjiang in China (the Chinese 'Burgess shale') show the presence of fish-like creatures in the early Cambrian (Shu et al., 1999)—between twenty and fifty million years earlier than was previously thought. In Prescott et al. (1999) we have reviewed evidence supporting the conservation, through evolution, of a basic vertebrate brain plan which may have been present in early jawless fish. Taken together, this evidence suggests that the first vertebrate nervous systems may be as ancient as any of those of the protostome bilaterian phyla. There is insufficient space to consider the many and varied forms of nervous system architecture seen in the protostome invertebrates. Instead, section 4 considers a number of aspects of the functional architecture of vertebrate brains that have implications for understanding how action selection occurs in vertebrates.

Finally, in section 5, we summarise our review of the evolution of action selection mechanisms in animal nervous systems and look for implications that could inspire the design of control architectures for autonomous robots. What exactly do we hope to find? The hope is that our study of comparative neurobiology will find evidence of what Dennett (1995) calls 'forced moves' in evolutionary design space, that is, outcomes imposed by strong task constraints; or 'good tricks'—robust and relatively general solutions to common problems, either of which could be usefully transferred into the design space for autonomous mobile beings.

2 Cnidarian nervous systems

Whilst the most primitive metazoans, the sponges, lack neurons and respond only to direct stimulation (usually with a very slow, spreading contraction), cnidarians have quite complex nervous systems, composed, principally, of distributed nerve nets, and show both internally generated rhythmic behaviour, and co-ordinated patterns of motor response to complex sensory stimuli.

The basic cnidarian nerve net is a two-dimensional network of neurons that has both a sensory and a motor capacity, and in which there is no distinction between axons and dendrites—nervous impulses therefore propagate in both directions between cells (Mackie, 1990). According to Horridge (1968), in the most primitive nerve nets "the spatial pattern is irrelevant, the connectivity pattern has no restrictions. […] any fibre is equivalent to any other in either growth or transmission" (p. 26).

The lack of intermediary forms of nervous system organisation between the aneural sponges and the cnidarian nerve net means that the evolutionary origin of nerve nets, and of nervous tissue in general, is shrouded in mystery. It seems likely, however, that neural conduction was preceded by more primitive forms of communication in which signals were propagated directly between neighbouring cells (indeed this form of non-neural communication exists alongside neural conduction in some cnidarians— Josephson, 1974). The evolution of the nerve net can then be understood as facilitating more rapid and more specific communication over longer distances, which would allow both quicker responses and increased functional diversification between different cell groups (Horridge, 1968; Mackie, 1990). Most of the neurophysiological features of more 'advanced' metazoan nervous systems are actually present at the cnidarian grade including multifunctional neurons, action potentials, synapses, and chemical neurotransmission. For Grimmelikhuijzen and Westfall (1995) the presence of such features shows cnidarians to be "near the main line" of evolution, and suggests that the study of their nervous systems will illuminate some of the properties of nervous systems ancestral to the higher metazoans.

The nervous systems of extant cnidarians are, in fact, more sophisticated than the above characterisation of simple nerve nets indicates. For instance, *Hydra*, one of the more primitive living cnidarians, has a variety of different neuronal cell-types, and while most belong to diffuse networks, some are found in localised, well-defined bundles that may have specific functional roles (Josephson & Mackie, 1965; Mackie, 1990). In other cnidarians, such as the hydrozoan jellyfish, parts of the nerve net are fused to form longitudinal or circular tracts that allow very fast signal conduction and can support fast attack, escape, or defense reactions. Many of the free-living cnidarians also possess light-sensitive and gravity-sensitive organs that allow behaviours such as orientation, sun compass navigation, and daily migration (see, e.g. Hamner, 1995); unfortunately the neural substrate that supports such behaviours remains poorly understood.

What is known about the functional architecture of cnidarian nervous systems? Horridge (1956; 1968) describes the decomposition of the nervous system of the jellyfish *Aurelia aurita* into two distinct components: a network of bipolar neurons that controls the symmetrical, pulsed contraction of the bell and enables the animal to swim; and a second more diffuse network, consisting largely of small multipolar neurons, that is spread across the body, tentacles, and margins of the animal, and coordinates localized feeding movements. These two systems, which are illustrated in figure 2 for the larva of *Aurelia aurita*, have relatively few interconnections and show clear evidence of independent operation. A similar functional subdivision of the nerve net into two or more parts has also been noted in a variety of other cnidarians such as sea anemones. This behavioural decomposition of control, with physically distinct circuits for feeding and movement, clearly shows an interesting similarity to that proposed for behaviour-based robots (see, e.g. Brooks, 1991).

The question arises, however, are alternative decompositions of the nervous system possible? Meech (1989) describes a jellyfish, *Aglantha digitale*, in which a single nerve net can carry two different types of action

potentials enabling either rapid escape swimming, or, slow rhythmic swimming for feeding. Similarly, the sea anemone *Actinia*, uses impulse patterns of different frequency to obtain distinct feeding and escape behaviours from a single nerve net (Mackie, 1990). Thus, there seems to be no strong requirement for a separate neural substrate for different classes of behaviour in these animals.

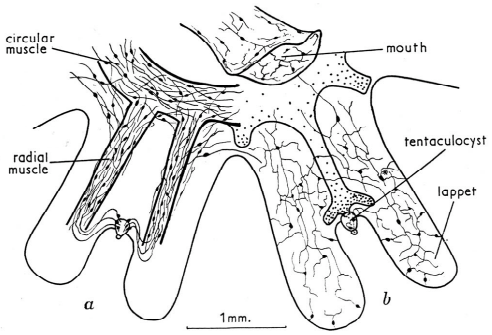


Figure 2. Nervous system of the ephyra larva of *Aurelia aurita*, showing, in two arms of the bell, a) the swimming network controlling the circular, and radial muscles; b) the diffuse nerve-net underlying feeding behaviour. A marginal ganglion is located at the base of each tentaculocyst. From Horridge (1956). with permission from the Company of Biologists Ltd.

The lack of centralised nervous system components in cnidarians also leads to some interesting and elegant solutions to the problem of generating an integrated global response. For instance, consider the fast escape behaviour of a jellyfish which can be triggered by contact at any point on the periphery of the animal. Since jellyfish swim by the synchronous, simultaneous contraction of the entire perimeter of the bell, the lack of centralised signalling presents an interesting control problem for which Mackie (1990) describes two contrasting solutions. One solution, seen in *Aglantha*, uses a giant axon with very fast conductance so that a single spike can circumnavigate the periphery in just a few milliseconds. An alternative and more remarkable solution, seen in the much larger species *Polyorchis*, involves a ring of neurons that carries action potentials that change shape as they circle the bell. Successive muscles groups respond to these changing shapes by contracting at shorter and shorter latencies, thus ensuring a uniform and synchronised contraction of the whole perimeter. This elegant solution appears to depend solely on membrane-level properties of the neurons involved (Spencer et al. 1989).

According to Horridge (1956) the two functionally distinct nerve nets of *Aurelia aurita* make contact with one another in neuron clusters termed the marginal ganglia. Each ganglion is part of the swimming network and is involved in the regular beat of the swimming contraction; it can also generate its own regular pulse if isolated from other parts of the network (thus showing an intrinsic rhythm generating capacity). Each ganglion is also in contact with the diffuse network that underlies the feeding response. Excitation in the diffuse network can inhibit the swimming rhythm or, in some cases, accelerate the rhythm. This evidence suggests the possibility of a hierarchical arrangement: pattern formation (the swimming beat) seems to be under the distributed control of multiple pace-maker systems, whilst the behaviour of this swimming network is under the modulatory control of the diffuse feeding network. If this is the case, then we could view this jellyfish nervous system as providing a natural example of a *subsumption architecture* (Brooks, 1986) composed of two distributed layers of control.

Cnidarian nervous systems demonstrate the ability of relatively simple nerve networks to support multiple behavioural modes, in some cases, using the same neural structures to generate two quite different patterns of activity. Whilst a likely physical substrate (the marginal ganglia) has been identified for the interaction between feeding and swimming in some jellyfish species there is no suggestion that these structures or pathways are exclusively involved in action selection. Although behavioural decomposition of function seems to be a probable cnidarian trait, decomposition involving specialised selection structures seems less likely. On the contrary, cnidarian nervous systems seem rather good preparations in which to study behavioural integration as a global, emergent property of the control system architecture.

3 Flatworm nervous systems

The phylum platyhelminthes comprises the free-living turbellarians and the parasitic flukes and tapeworms. The focus here will be on the turbellarians as the consensus in modern zoology is that these animals provide a better indication of the ancestral characteristics of the phylum. In the evolution of bilateral animals a critical development was the appearance of a central nervous system organised around a massed concentration of nerve cells called the *cephalic ganglion*—the 'archaic brain'. In flatworms we find the simplest living animals that possess this form of nervous system architecture (Reuter, 1989).

Flatworms are bilaterally symmetric having distinct anterior and posterior ends, and dorsal (upper) and ventral (lower) surfaces. Sensory systems are distributed symmetrically between the left and right sides of the body, but together with the nervous system often show a concentration, termed *cephalization*, towards the anterior end of the body. The free-living turbellarians range in size from a few millimetres to tens of centimetres. They are found in aquatic environments or moist terrestrial environments where most pursue a predatory or scavenging life-style requiring a repertoire of reasonably complex behaviours. Turbellarian nervous systems appear in a bewildering variety of different configurations, none of which can necessarily be considered primitive (Reuter, 1989). Typically, there are three to five pairs of major nerve cords connecting with the cephalic ganglion. These cords are interlinked by circular commissures (bands of nerve fibres), which themselves make connections with networks (plexuses) of nerves underlying muscular and/or epithelial tissue. The cell bodies of sensory neurons are found near the periphery while those of motor neurons and interneurons are distributed throughout the nerve cords and the brain. The concentration of nerve cells into cords, fibres, and ganglia distinguishes this type of central nervous system from the nerve nets of the cnidaria.

Our discussion of the functional architecture of the flatworm nervous system follows the research of Gruber and Ewer (Gruber & Ewer, 1962) and of Koopowitz and co-workers (reviewed in Koopowitz & Keenan, 1982) which has focused on the role of the brain in marine polyclad turbellaria.

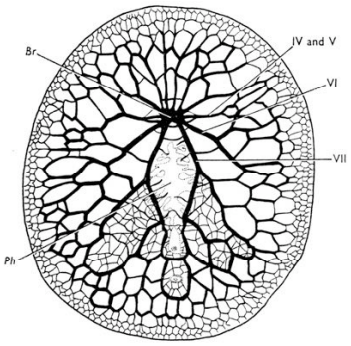


Figure 3. Nervous system of the turbellarian *Planocera gilchristi*, showing the brain (Br), pharynx (Ph), and major nerve cords (1V-VII). The finest granularity of nerve fibres is only shown in the central areas around the pharynx. From Gruber and Ewer (1962) with permission from the Company of Biologists Ltd.

Gruber and Ewer studied the effect of brain removal on the behaviour of the polyclad *Planocera gilchristi*, whose nervous system is pictured in figure 3. Planocera usually moves by swimming or crawling along the substrate. Swimming involves the generation of a transverse wave that moves backwards along the length of the body, while crawling involves a regular alternating extension of the two sides of the body. Following brain removal, Gruber and Ewer reported that components of both normal swimming and normal crawling were present in decerebrate animals but that these were never integrated into the normal sequences—the overall movement of the animal was irregular and uncoordinated. Similarly, decerebrate animals lacked a normal rapid righting response when placed in an inverted position, although they could eventually right themselves by making writhing and twisting movements. These animals also failed to display the normal retraction response to mechanical stimulation, again responding with an uncoordinated writhing.

Gruber and Ewer also describe the effects of decerebration on the feeding behaviour of *Planocera*. This behaviour was the subject of further detailed investigation by Koopowitz who went on to examine decerebrate feeding in another marine polyclad—*Notoplana*. The behaviour of *Notoplana* will be described here as it is typical of the general pattern of results obtained with these animals.

In the intact polyclad worm presentation of a food item near to its posterior margin will cause it to extend a nearby portion of that margin and use this to take hold of the food. The animal will then rotate the anterior part of its body on that side, until the anterior margin comes into contact with the food. The posterior margin subsequently loosens its grip allowing the anterior edge to manipulate the food into the mouth. This sequence of behaviour is shown in figure 4a. When fed with large food items (dead shrimps) the animal becomes satiated after a few food presentations and the feeding response ceases.

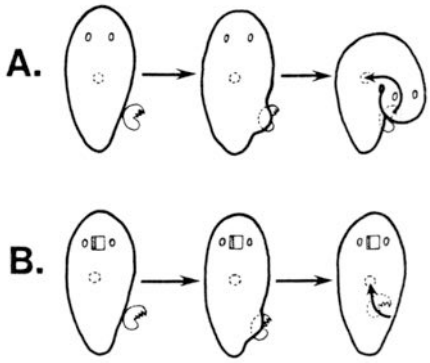


Figure 4. Feeding behaviour of the polyclad *Notoplana*. A. In the intact animal contact with an item food at the posterior margin causes a whole body response in which the animal turns and grabs the food with its anterior edge before passing it to the central, ventrally located mouth. B. In the decerebrate animal, a 'local feeding response' causes food to be passed directly to the mouth. From Koopowitz and Keenan (1982) with permission from Elsevier Science.

In the decerebrate animal, in contrast, the body turn to bring the anterior margin into contact with the food is never observed. Instead, the animal performs a 'local feeding response' in which it gradually moves the food directly to its mouth via the underside of its body (figure 4b). In addition to the lack of a coordinated 'whole body' feeding response the decerebrate animals show no satiety and will continue passing food items towards the mouth even once the gut is completely full. Control experiments in which the brain remains intact but the main posterior nerves on one side of the body are severed, show feeding behaviour characteristic of the normal animal on the intact side, and that characteristic of the decerebrate on the cut side.

Overall, these experiments on decerebrate polyclad behaviour demonstrate the role of the brain in regulating local reflexive actions whose neural substrate is located in the periphery of the animal. In the case of crawling and swimming, the brain orders the temporal sequence of local activity in different marginal areas of body. In the case of feeding, the brain holds the 'local feeding response' under inhibitory control whilst enabling actions involved in the 'whole body' feeding response.

The centralized coordination of behaviour seen in the polyclad stands in interesting contrast to the distributed nature of control noted in the cnidaria. What evolutionary pressures may have brought about such a significant change in the functional organization of nervous systems? Koopowitz and Keenan (1982) contrast two possible explanations for the evolution of the first brains. The first possibility is that the brain is one of several consequences of the process of cephalization—the aggregation of sensory systems in the anterior portion of the animal. According to this explanation, the co-ordination of peripheral mechanisms becomes focused in the brain in order to place it closer to the principle sources of afferent stimulation. This view also makes the primary role of the archaic brain one of response initiation. The alternative view, favoured by Koopowitz and Keenan, is

based on the observation that although all polyclads have brains, only a few show significant cephalization. Instead, the origin of the brain could be attributable to a more fundamental change in the body plan of the organism—the evolution of bilateral symmetry:

"We consider that the development of bilateral symmetry, rather than cephalization, was the prime feature that necessitated the evolution of the brain. Bilateral symmetry required that the righthand side know what was happening on the left, and vice versa. In effect, with the advent of bilateral symmetry, the evolution of the brain was necessary for the coordination of disparate peripherally-based reflexes. This was of prime importance in preventing the two sides from engaging in contradictory activities." (Koopowitz and Keenan, 1982, p. 78)

From the perspective of this paper, this interesting proposal might be paraphrased as the hypothesis that the brain first evolved as a centralised mechanism for action selection.

Koopowitz and co-workers also describe further experiments in which half of the polyclad brain is excised, and the severed cephalic nerve cords allowed to regrow, re-establishing appropriate functional connections with the remaining half-brain. In other experiments the brain of one animal is transferred in its entirety into another's body and once again re-exerts many of its original behavioural controls over the periphery. Finally, brain-control returns even if the brain is re-inserted upside-down or rotated 180° (try doing this with a robot's CPU!). This robustness of function is particularly remarkable given that the brain is clearly much more than a relay station between the two halves the animal, but instead plays an integrative role in selecting appropriate patterns of peripheral motor acts.

4 Vertebrate nervous systems

The evolution of the vertebrate nervous system is a critical unsolved problem in evolutionary neurobiology. Vertebrates belong to the phylum chordata whose members all possess, at some stage in their development, a single, hollow nerve cord, called the neural tube, which runs most of the length of the longitudinal body axis. Unfortunately, all living protochordates (that is, animals of the chordate phylum that are *not* vertebrates) have relatively simple nervous systems, and only one species, *Branchiostoma* (previously known as *Amphioxus*), has a nervous system that could resemble a transitional stage between ancestral chordate and vertebrate. Branchiostoma shows elaborations at the anterior end of the neural tube that may be homologous to some regions of the vertebrate brain (Lacalli, 1996); however the 'brain' of Branchiostoma is tiny, it sensory systems primitive, and its behaviour very simplified compared with that of living vertebrates. In the modern fauna, the most primitive vertebrate characteristics are found amongst the jawless fish (*Agnatha*). Examination of these animals has shown the same gross morphological divisions of the nervous system—spinal cord, hindbrain, midbrain, and forebrain—as are present in other vertebrate classes. Indeed, impressions of these structures have also been found in the fossilized endocasts (casts from the inside of fossil skulls) of ancient agnathans. This evidence suggests that a basic 'ground plan' for the nervous system is shared by all living vertebrate classes, and possibly by all ancestral vertebrates (see Prescott et al, 1999).

The substrate for action selection in a control architecture as complex as the vertebrate nervous system is likely to involve many different mechanisms and structures. The following brief review is by no means exhaustive but considers a few promising candidates. Beyond the mechanism identified here, selection as the emergent consequence of interactions between circuits with wider functional roles may also play an important role.

Conflict resolution for clean escape

One of the requirements for effective action selection is timely, sometimes very rapid, decision making. Transmission and response times in neural tissue are not negligible so for urgent tasks it is important to ensure that time is not lost resolving conflicts with competing behaviours. Indeed, there is evidence to suggest, that for tasks such as defensive escape, special circuitry may have evolved in the vertebrate nervous system to provide a very fast override of the competition. The giant *Mauthner* cells (M-cells) found in the brain-stem of most fish and some amphibians provide an example of this function. M-cells are known to be involved in the 'C-start' escape maneuver—the primary behaviour used by many species of fish to avoid hazards such as predation. Eaton, Hofve, and Fetcho (1995) have argued that the principle role of the M-cell in the brainstem escape circuit may *not* be to initiate the C-start but to *suppress competing behaviours*. This conclusion is supported by evidence that removal of the M-cells does not disable the C-start or have a marked effect on the strength or latency of the response. Instead, the fast conduction of the Mauthner giant axon (one of the largest in the vertebrates) may be crucial in ensuring that contradictory signals, that could otherwise result in fatal errors, do not influence motor output mechanisms. Conservation of brain-stem organization across the vertebrate classes suggests that homologous mechanisms may play a similar role in the escape behaviours of other vertebrates.

Fixed priority mechanisms

Many studies of the role of the vertebrate brain in behavioural integration suggest that the resolution of conflict problems between the different levels of the neuraxis (spinal cord, hind-brain, mid-brain, etc.) may be determined by fixed-priority, vertical links. For instance, in (Prescott et al., 1999) we have reviewed evidence that the vertebrate defense system can be viewed as a set of dissociable layers in which higher levels can suppress or modulate the outputs of lower levels (using mechanisms somewhat similar to the inhibition and suppression operators employed in the subsumption architecture). Fixed-priority mechanisms cannot, however, capture the versatility of behaviour switching observed between the different behaviour systems (defense, feeding, reproduction, etc.) found in adult vertebrates. Since dominance relationships between behaviour systems can fluctuate dramatically with changing circumstances more flexible forms of selection are required than can be determined by hard-wiring.

Reciprocal inhibition

A specific form of neural connectivity, often associated with action selection, is mutual or reciprocal inhibition (RI). In networks with recurrent reciprocal inhibition two or more sub-systems are connected such that each one has an inhibitory link to every other. Such circuits make effective action selection mechanisms since the most strongly activated sub-system will receive less total inhibition than any of the others; and the recurrent connectivity of the system results in positive feedback that rapidly maximises the activity of this 'winner' relative to all the other 'losing' sub-systems. RI connectivity has been identified in many different areas of the vertebrate brain (Windhorst, 1996) and could play a role in conflict resolution at multiple levels of the nervous system (Gallistel, 1980). One likely locus for selection via RI is the *superior colliculus* in which within-nucleus long-range inhibitory and short-range excitatory connections may co-operate to select a single target for visual orienting out of several available candidates. This possibility is discussed in our companion paper (Chambers, Gurney, & Prescott, this volume).

Centralised selection mechanisms

Snaith and Holland (1990) have contrasted a distributed action selection based on RI with a system that employs a specialized, central switching device. They note that to arbitrate between n competitors, an RI system with full connectivity requires n(n-1) connections, while adding a new competitor requires a further 2n connections. In contrast, a system using a central switch requires only 2 connections per competitor (to and from the switch) resulting in 2n connections in all. Adding a further unit requires only 2 additional connections. On this comparison, a central switching device clearly provides a significant advantage in terms of economy of connections costs. Ringo (1991) has pointed out that geometrical factors place important limits on the degree of network interconnectivity within the brain. In particular, larger brains cannot support the same degree of connectivity as smaller ones—significant increases in brain size (as have been seen in vertebrate evolution) must inevitably be accompanied by decreased connectivity between non-neighbouring brain areas. Since functional units in different parts of the brain will often be in competition for the same motor resources, the requirement of lower connectivity with increased brain-size therefore strongly favours selection architectures with lower connectional costs.

We have proposed (Prescott et al., 1999; Redgrave, Prescott, & Gurney, 1999) that a group of functionally related structures in the vertebrate fore- and mid- brain, called the *basal ganglia*, appear to be suitably connected and configured to serve as an array of specialized central switching devices that could provide effective conflict resolution with economical interconnectivity. We have developed a number of computational and robotic models of the basal ganglia from this perspective (see Prescott et al, in press; Gurney, Prescott, Wickens, & Redgrave, 2004, for reviews), including one described in our companion paper (Chambers et al., this volume) that specifically considers the interaction between basal ganglia selection mechanisms and more localised selection circuits in the superior colliculus.

Studies of infants rats in whom the basal ganglia are not yet developed (see, e.g. Berridge, 1994), and in animals in which the forebrain has been removed, indicate that, below the basal ganglia, there is a brainstem substrate for selection that, at the very least, can provide appropriate behaviour switching while the adult architecture is developing or when it is damaged or incapacitated. One likely locus for this mechanism is in the medial core of the brainstem *reticular formation* (RF). Our group is also involved in investigating computational models of selection in this structure (Humphries, Gurney, & Prescott, In press) the latest of which is described in a second companion paper (Humphries, Gurney, & Prescott, this volume).

Both the basal ganglia and the RF medial core lie in central positions along the main neuraxis (see figure 5). They have been described collectively as forming the brain's 'centrencephalic core', and identified by a number of neurobiologists as playing a key role in the integration of behaviour (see Prescott et al., 1999 for review). In the intact adult brain, it is likely that both systems co-operate, in some unknown manner, to determine what form of behaviour is expressed at a given time.

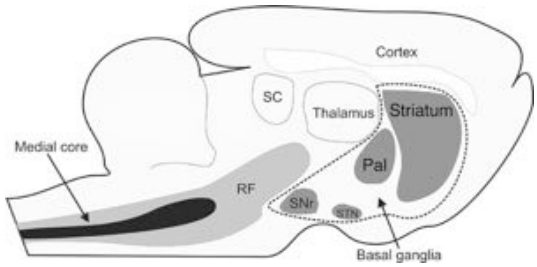


Figure 5. Systems for action selection in the vertebrate brain: A saggital slice through the rat brain illustrating the locations of medial core of the reticular formation (RF), the basal ganglia (Striatum, Pal—Pallidum, SNr—Substantia Nigra, STN—subthalamic nucleus), and the superior colliculus (SC).

6 Conclusions: 'forced moves' and 'good tricks' in the evolution of action selection

We have provided a brief review of the neural substrate of action selection in a number of living animal groups. Our review has been limited to discussing neural circuits although there is good evidence that other mechanisms, such as the endocrine system, can play an important role in action selection (see, e.g. Barrington, 1967; Brooks, 1994). Despite these limitations, we believe that a number of conclusions can be drawn from the findings we have surveyed.

First, the investigation of cnidarian nervous systems shows that many forms of behavioural integration can be achieved in complex multi-celled animals in the relative absence of centralised nervous system structures. The elegance of these natural solutions is only just beginning to be matched by those developed for distributed robot control systems. We suspect that the study of cnidarian nervous systems and behaviour could provide some 'good tricks' for the design of future 'minimalist' mobile robots.

Second, our review of flatworm nervous systems suggests that the evolution of centralised selection mechanisms in the archaic brain may have been a 'forced move' required to maintain behavioural coherence in a bilaterally-organised animal. It seems likely that the design of artificial control systems could benefit from the use of similar centralised conflict resolution systems because of the advantages that this form of modularity can confer (see Prescott et al., 1999; Bryson, 2000).

Finally, our review of the neural substrate of action selection in vertebrates has identified a number of candidate mechanisms that may be instantiated in their neural circuitry. This evidence suggests the existence of multiple substrates for action selection in the vertebrate nervous system. A key proposal is that vertebrates exploit specialized selection circuitry found in groups of centralised brain structures—the medial core of the reticular formation and the basal ganglia. It seems possible that the connectional economy of this centralised design, which can act to resolve competitions between functional sub-systems distributed widely in the brain, may be one the reasons that the vertebrate nervous system has scaled successfully with the evolution of animals of larger brain and body size. The design of control systems for robots with multiple actuator sub-systems should benefit from a better understanding of how these different elements of the vertebrate nervous system co-operate to maintain behavioural coherence.

References

Barrington, E. J. W. (1967). *Invertebrate structure and function*. Sunbury-on-Thames, UK: Nelson.

Berridge, K. C. (1994). The development of action patterns. In J. A. Hogan & J. J. Bolhuis (Eds.), *Causal Mechanisms of Behavioural Development*. Cambridge, UK: CUP.

Brooks, R. A. (1986). A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation, RA-2*, 14-23.

Brooks, R. A. (1991). New approaches to robotics. *Science, 253*, 1227-1232.

Brooks, R. A. (1994). *Coherent behaviour from many adaptive processes.* From Animals to Animats 3: Proceedings of the Third Int. Conf. on the Simulation of Adaptive Behaviour, Brighton, UK.

Bryson, J. (2000). Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental and Theoretical Artificial Intelligence, 12*(2), 165-189.

Chambers, J., Gurney, K. N., & Prescott, T. J. (Submitted to the MNAS workshop). *Mechanisms of choice in the primate brain: a quick look at positive feedback.*

Dennett, D. (1995). *Darwin's Dangerous Idea*. London: Penguin.

Eaton, R. C., Hofve, J. C., & Fetcho, J. R. (1995). Beating the competition - the reliability hypothesis for Mauthner axon size. *Brain Behavior and Evol., 45*(4), 183-194.

Gabor Miklos, G. L., Campbell, K. S. W., & Kankel, D. R. (1994). The rapid emergence of bio-electronic novelty, neuronal architectures, and organismal performance. In R. J. Greenspan (Ed.), *Flexibility and Constraint in Behavioural systems*: John Wiley and Sons.

Grimmelikhuijzen, C. J. P., & LA. Westfall, L. A. (1995). The nervous systems of Cnidarians. In O. Breidbach & W. Kutsch (Eds.), *The Nervous Systems of Invertebrates An Evolutionary and Comparative Approach* (pp. 7-24). Basel, Switzerland: Birkhauser Verlag.

Gruber, S. A., & Ewer, D., W. (1962). Observations on the myo-neural physiology of the polyclad. *Planocera gilchristi. Journal of Experimental BIology, **39***, 459-477.

Gurney, K., Prescott, T. J., Wickens, J., & Redgrave, P. (2004). Computational models of the basal ganglia: from membranes to robots. *Trends Neurosci., 27*, 453-459.

Hamner, W. M. (1995). Sensory ecology of scyphomedusae. *Marine and Freshwater Behaviour and Physiology, 26*(2-4), 101-118.

Horridge, A. (1956). The nervous system of the ephyra larva of *Aurellia aurita*. *Quarterly Journal of Microscopical Science, 97*, 59-74.

Horridge, G. A. (1968). The origins of the nervous system. In G. H. Bourne (Ed.), *The Structure and Function of Nervous Tissue* (Vol. 1, pp. 1-31). New York: Academic Press.

Humphries, M. D., Gurney, K. N., & Prescott, T. J. (In press). Is there an integrative center in the vertebrate brain-stem? A robotic evaluation of a model of the reticular formation viewed as an action selection device. *Adaptive Behavior*.

Humphries, M. D., Gurney, K. N., & Prescott, T. J. (Submitted to the MNAS workshop). *Action selection in a macroscopic model of the brainstem reticular formation.*

Josephson, R. K. (1974). Cnidarian neurobiology. In L. Muscatine & H. M. Lenhoff (Eds.), *Coelenterate biology: reviews and new perspectives* (pp. 245-273). New York: Academic Press.

Josephson, R. K., & Mackie, G. O. (1965). Multiple pacemakers and the behaviour of the hydroid *Tubularia*. *Journal of Experimental Biology, 43*, 293-332.

Koopowitz, H., & Keenan, L. (1982). The primitive brains of platyhelminthes. *Trends in Neurosciences, 5*(3), 77-79.

Lacalli, T. C. (1996). Frontal eye circuitry, rostral sensory pathways and brain organization in amphioxus larvae - evidence from 3d reconstructions. *Philosophical Transactions Of the Royal Society Of London Series B- Biological Sciences, 351*(1337), 243-263.

Mackie, G. O. (1990). The elementary nervous-system revisited. *American Zoologist, 30*(4), 907-920.

Maes, P. (1995). Modelling adaptive autonomous agents. In C. G. Langton (Ed.), *Artificial Life: An Overview*. Cambridge, MA: MIT Press.

McFarland, D. (1989). *Problems of Animal Behaviour*. Harlow, UK: Longman.

Meech, R. W. (1989). The electrophysiology of swimming in the jellyfish *Aglantha digitale*. In P. A. V. Anderson (Ed.), *Evolution of the first nervous systems*. New York: Plenum Press.

Prescott, T. J., & Ibbotson, C. (1997). A robot trace-maker: modeling the fossil evidence of early invertebrate behavior. *Artificial Life, 3*, 289-306.

Prescott, T. J., Redgrave, P., & Gurney, K. N. (1999). Layered control architectures in robots and vertebrates. *Adaptive Behavior, 7*(1), 99-127.

Prescott, T. J., Montes Gonzalez, F., Gurney, K. N., Humphries, M. D., & Redgrave, P. (In press). A robot model of the basal ganglia: behaviour and intrinsic processing. *Neural Networks*.

Raff, R. A. (1996). *The Shape of Life: Genes, Development and the Evolution of Animal Form*. Chicago: Chicago University Press.

Raup, D. M., & Seilacher, A. (1969). Fossil foraging behaviour: computer simulation. *Science, 166*, 994-995.

Redgrave, P., Prescott, T., & Gurney, K. N. (1999). The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience, 89*, 1009-1023.

Reuter, M. (1989). From innovation to integration: Trends of the integrative systems in microturbellerians. In M. K. S. Gustafsson & M. Reuter (Eds.), *The Early Brain* (pp. 161-178). Abo, Finland, Abo Academy Press.

Ringo, J. L. (1991). Neuronal interconnection as a function of brain size. *Brain, behaviour, and evolution, 38*, 1-6.

Shu, D.-G., Luo, H.-L., et al. (1999). Lower Cambrain vertebrates from south China. *Nature, **402***, 42-46.

Snaith, S., & Holland, O. (1990). *An investigation of two mediation strategies suitable for behavioural control in animals and animats.* From Animals to Animats 1.

Spencer, A. N., Przysiezniak, J., & Acosta-Urquidi, J. et al. (1989). Presynaptic spike-broadening reduces junctional potential amplitude. *Nature, 340*, 636-638.

Windhorst, U. (1996). On the role of recurrent inhibitory feedback in motor control. *Prog. In Neurobiology, 49*(6), 517-587.

Mechanisms of choice in the primate brain: a quick look at positive feedback

JM Chambers, K Gurney, M Humphries, A Prescott
Department of Psychology, University of Sheffield
Western Bank, Sheffield, S10 2TP
j.m.chambers@sheffield.ac.uk

Abstract

We consider the possibility that positive feedback loops are exploited by the brain in determining which action to perform at any given moment. We emphasise the need for, and requirements of, a controller that can exploit the potential benefits, and overcome the inherent pitfalls of using positive feedback for selection. We present the vertebrate basal ganglia as one possible solution to this control problem, and focus on basal ganglia involvement in the oculomotor system of the primate brain, presenting it as an example of how positive feedback and competitive dynamics are used synergistically to bring about changes in gaze. Finally we strengthen the case for involvement of positive feedback mechanisms in reflexive gaze control by demonstrating that a computational model of the oculomotor system is able to reproduce eye movement abnormalities present in sufferers of Parkinson's disease - a disease that affects the basal ganglia, and consequently the control of positive feedback.

1 Introduction

Humans make approximately 3 eye movements every second. Some of these are made deliberately, for instance when reading, while others are made in response to external events. In a complex environment there are apt to be a countless number of objects vying for attention. How then does the brain determines which of these is worthy of further scrutiny, and how does it ensure that the eyes are guided to that object accurately, without interference from competing targets?

The answer to these questions may lie in the discovery of anatomical links between the oculomotor system and the basal ganglia (BG), a set of deep brain nuclei that are implicated in decision making [Hikosaka *et al*., 2000; Redgrave *et al*., 1999]. The work reported here seeks to explore the nature of this link through the use of computational models. In particular, we focus on an explanation for the neural activity recorded in reactive saccade tasks, and how this relates to certain reaction time phenomena observed in Parkinson's Disease (PD) patients.

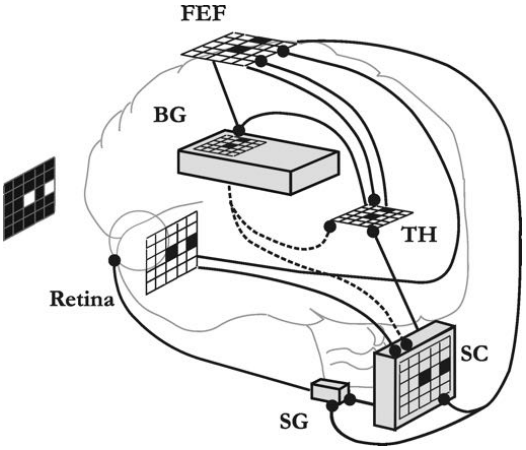


Figure 1: Brain areas forming the reactive oculomotor system. SC - superior colliculus; SG - saccadic generator; TH - thalamus; FEF - frontal eye fields; BG - basal ganglia. Solid and dashed lines denote excitatory and inhibitory projections respectively.

1.1 Oculomotor abnormalities in Parkinson's disease

PD is a degenerative disease characterised by the death of midbrain neurons that produce the neuro-modulator dopamine (DA). The input nucleus of the BG, the striatum, is a major target of these DA cells, and consequently their death causes a loss of modulatory control over the BG. PD patients show a number of characteristic abnormalities in saccadic control. These include hypometric saccades (undershooting the target), decreased saccadic velocity, and failure to generate saccades (akinesia) (see [Kennard and Lueck, 1989] for review). Interestingly, in some experimental paradigms, PD patients also show a 'paradoxical' reduction in response time (RT; the length of time between target onset and saccade initiation). The design and results of one such experiment [Briand *et al*., 1999] are shown in figure 2. The experiment demonstrates a small RT advantage for PD (of $\sim$10 ms), and a reduction in saccade amplitude ($\sim$5%; see [Briand *et al*., 2001], for an experiment that yields a more significant result). Because PD is a disease that almost exclusively affects the BG, we hope that by attempting to explain the result of Briand et

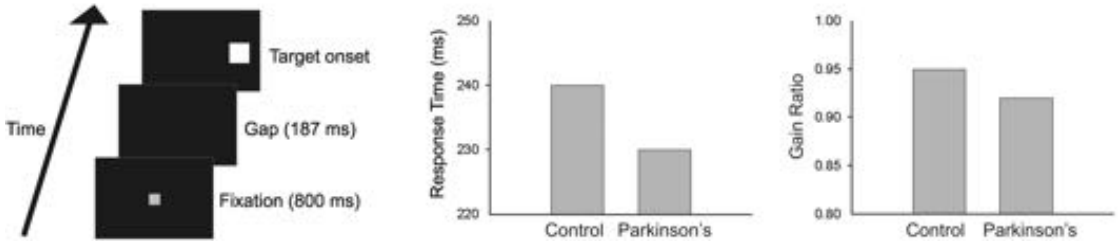


Figure 2: The experimental paradigm used by Briand et al., [1999] to test reactive saccades in PD patients. Subjects fixate a central stimulus for 800 ms. This is extinguished, and 187 ms later one of two possible target stimuli is illuminated to which the subject makes a saccade. RTs are measured from the time of target onset to the time of saccade onset. Gain ratio is the ratio of final eye displacement to actual displacement required to centre the gaze on the target.

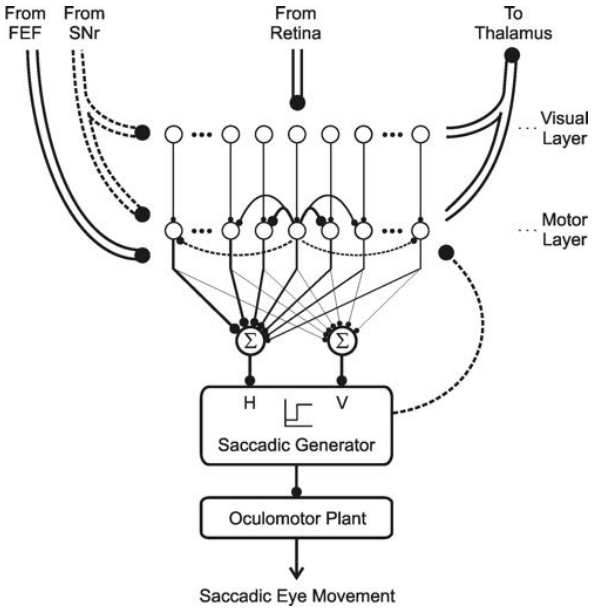


Figure 3: A model of the SC based on Arai et al.[1994]. Solid and dashed lines denote excitatory and inhibitory projections respectively. Double-lines denote topographic projections. See text for description.

al., we will gain further insight into the role of the BG within the oculomotor system and, more generally, as the neural substrate for decision making.

1.2 The oculomotor system

The superior colliculus and saccadic generator

Retinal ganglion cells project directly to the superior colliculus (SC; [Schiller and Malpeli, 1977]), a multi-layered, midbrain structure, that preserves the spatial organisation of its retinal input. Figure 3 shows the basic connectivity of the SC as implemented in the model of Arai et al. [1994] (hereafter referred to as the Arai model) which we have incorporated into our own large-scale model (discussed in methods section). The superficial layer of the SC relays its phasic retinal input to deeper motor layers, which in turn, send excitatory

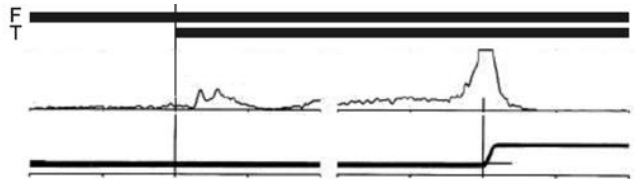


Figure 4: A typical VM response. Bars marked F and T denote the presence of the fixation and target stimuli respectively. Middle trace shows mean response of a group of VM cells recorded from the SC motor layer of a monkey. Bottom trace denotes position of eye.

projections to a set of brainstem nuclei, collectively known as the saccadic generator (SG) circuits, which provide closed-loop control of the eye muscles [Sparks, 2002].

The inner workings of the SG are beyond the scope of this paper, however, one important detail of SG operation is key to understanding later discussions. Models of the SG invariably incorporate a class of cell known as an omni-pause-neuron (OPN), that are thought to actively inhibit the neurons which drive changes in eye position. In a recent SG model proposed by Gancarz et al. [1998] (hereafter referred to as the Gancarz model), the saccade command that the SG receives from the SC and FEF, is responsible for inhibiting the OPNs, so that saccades will only be initiated if the saccade command is of sufficient magnitude, and as such the OPNs provide a threshold for action (as indicated by the step icon in figure 3).

The frontal eye field

Another important source of input to the SC comes from the frontal eye field (FEF), an area of the frontal lobes implicated in saccade generation. The FEF receive (among other sources) a strong input from the posterior cortices that comprise the 'where' pathway of visual processing. The nature of the processing that takes place in the posterior cortices is not important for our purposes (as alluded to by the direct connection between the retina and the FEF in figure 1), other than to say, that it preserves a retinotopic organisation, and that it displays tonic activation when a visual stimulus is present on the retina. In addition to projecting to the SC, the FEF also project directly to the SG so that a subject with a SC lesion is

still able to generate saccades.

The visuo-motor response

Electrophysiological studies with primates have revealed that neurons in the SC and FEF display very similar patterns of activity during oculomotor tasks. One common class - the visuomotor (VM) response - is observed when a stimulus suddenly appears in peripheral vision and subsequently forms the target of a saccade. VM cells display a bimodal activity profile, in which the first peak (visual) is a phasic response locked to the time of stimulus onset, and the second peak (motor) is a phasic response that coincides with the time of saccade onset (Figure 4; [Munoz and Wurtz, 1995]). In their landmark study, Hanes & Schall [1996] demonstrated that saccadic RT is determined by the rate at which FEF motor activity grows towards a threshold firing rate, consistent with psychological models of decision making [Ratcliff, 1978].

Positive Feedback in the Oculomotor System

Given that the motor component of the VM response is critical in determining RT, it is interesting to consider what causes it. Arai et al. [1994] suggest that the build up is generated by local excitatory loops within the SC motor layers and triggered by the BG (Figure 3). The substantia nigra pars reticulata (SNr) - one of the output nuclei of the BG - provides strong tonic inhibition to the SC motor layer, but this is known to pause just prior to saccade initiation [Hikosaka *et al*., 2000]. The Arai model shows that this disinhibition can cause residual visual activity in the SC motor layers to be amplified by local (SC-SC) positive feedback. Inspection of figure 1 reveals that the oculomotor system contains at least two additional positive feedback loops (PFBLs): SC-TH-FEF-SC, and FEF-TH-FEF (TH = thalamus) [Sommer and Wurtz, 2004; Haber and McFarland, 2001]. The pause in BG output can affect activity in all three, as in addition to targeting the SC, the SNr also projects to TH. It is therefore likely that the buildup of motor activity observed in the SC and FEF is in fact produced by the combined effect of all three PFBLs.

Both the FEF and TH project back to the BG [Hikosaka *et al*., 2000; Harting *et al*., 2001] with retinotopic projections, so that activity in FEF, TH and SC, is both affected by, and able to affect BG output. The striatum - a BG input nucleus - sends an inhibitory projection to the SNr (Figure 5), so that activation of striatal neurons can cause a pause in SNr firing. Like much of the oculomotor system, the BG have a retinotopic organisation, and recent evidence suggests that the projection from SNr to SC preserves this mapping [Basso and Wurtz, 2002] so that localised input to the BG may be able to cause localised disinhibition in the SC, meaning that the BG output determines not just *when*, but also *where* saccade-related activity is able to buildup within the SC motor-map.

1.3 Competition in the oculomotor system

In addition to the retinal and FEF input shown in figure 1, the SC also receives excitatory input from several visual, auditory, and somatosensory areas of cortex, so that saccades can be triggered by processed visual features, localised noises, or physical contact with the body [Stein, 1993]. Clearly, for an animal operating in a complex environment, there will be moments when the SC's multi-modal inputs are sending conflicting commands. Under the scheme described so far, it would seem that visual input to the BG leads to the inevitable disinhibition of the SC motor layer, and a saccade towards the stimulus causing it. Clearly this cannot be the case, somewhere in the oculomotor circuit, a decision is being taken as to which location should be attended to, be it the currently fixated point or any other.

Reciprocal inhibition (RI) is a form of neural connectivity that is often associated with action selection, and found throughout the vertebrate brain [Windhorst, 1996]. RI gives rise to winner-take-all (WTA) dynamics, as the most active neural population is able to silence its competitors. If all coordinates in the oculomotor system are to compete with each other via RI, then each part of the retinotopic map in a nucleus, must be connected to every other part. While there is evidence for RI connectivity in the oculomotor cortex, BG, and SC [Windhorst, 1996; Munoz and Istvan, 1998; Meredith and Ramoa, 1998], it is unlikely that this is sufficiently long-range to enable competition between all coordinates.

Gurney et al. [2001] suggest that the BG may contain a type of feed-forward selection circuit that differs from RI. Figure 5 shows their computational model (hereafter referred to as the Gurney model), and provides a description of how intrinsic BG processing achieves signal selection. The extent to which a channel is selected is determined by the difference between its own activity and the sum of all channel activity. The calculation takes place in SNr, where diffuse excitatory input from the sub-thalamic nucleus (STN) effectively provides the sum of channel activity, and focused inhibitory input from D1 striatal cells provides a measure of individual channel activity.

The diffuse STN projection allows inter-channel communication, so that input to a given BG channel acts to raise the level of inhibition outputted from all other channels. Thus, the growth rate of motor activity in a BG controlled PFBL, will depend not only on the sensory input driving it, but also on the activity in other BG controlled loops. So that for instance, in the experiment used by Briand et al. [1999], activity in a loop corresponding to the fixation coordinate, will affect activity in a loop corresponding to the target coordinate.

The Gurney model identifies a control pathway through the BG that modulates STN activity to keep selection optimal. The control pathway incorporates the D2 striatal neurons, as opposed to the D1 population involved in selection. DA makes D1 cells more excitable, while making D2 cells less so. Consequently, a disruption in the level of tonic DA received by the striatum, affects the balance between the selection and control pathways. The Gurney model predicts that low DA levels (as present in PD) will cause sub-optimal selection, with incomplete disinhibition of the winning channel.

This prediction, when combined with the likely role of BG disinhibition in generating oculomotor buildup activity, suggests that the RT differences between controls and PD patients might result from abnormal competitive dynamics within the oculomotor system. Before considering how this might work using our large-scale model of the oculomotor system, we first familiarise the reader with the properties of positive feedback under inhibitory control.

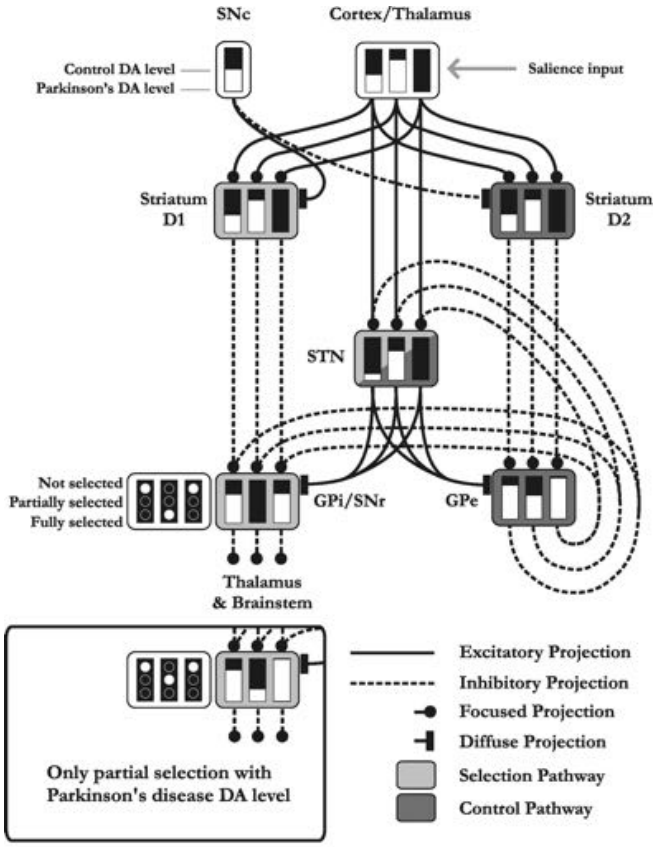


Figure 5: The intrinsic BG model of Gurney et al.[2001b], assumes that duplicate salience input is sent to the sub-thalamic nucleus (STN) and striatum, which is further sub-divided in two groups of cells classified by the type of dopamine (DA) receptor they express (D1 and D2). The globus pallidus internal segment (GPi) and substantia nigra pars reticulata (SNr) - which together form the output nuclei of the BG - send inhibitory projections back to thalamus and to motor nuclei in the brainstem (e.g., the SC). Spontaneous, tonic activity in the STN guarantees that this output is active by default, so that all motor systems are blocked. Gurney et al., identify two separate functional pathways within the BG. The selection pathway is responsible for disinhibiting salient actions: salience input to a channel activates D1, which then inhibits GPi/SNr thus silencing inhibitory output in the channel. The diffuse projection from STN to GPi/SNr means that all channels receive an increased excitatory drive. This is offset in the most active channel by the inhibitory input from D1, but goes unchecked in less active channels thus acting to block unwanted actions. The control pathway defined by Gurney et al., incorporates the globus pallidus external segment (GPe), and provides capacity-scaling by ensuring that STN activity does not become excessively high when multiple channels have non-zero salience, thus assuring full disinhibition of the winning channel irrespective of the number of competing channels. Because the striatal input to the control and selection pathways utilise different DA receptors, changes in tonic DA levels affect them differentially. Consequently, when DA is reduced to PD-like levels, the balance between the two pathways is disturbed resulting in residual inhibition on the selected channel (inset).

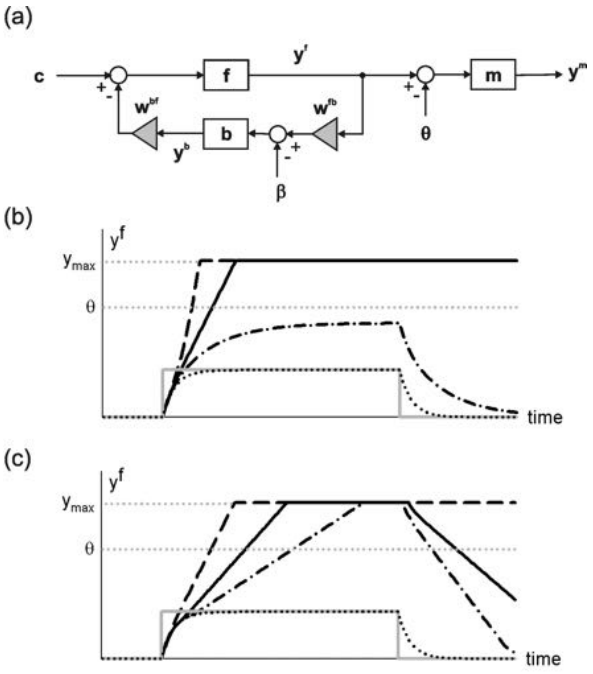


Figure 6: a) A simple behavioural control system incorporating positive feedback. b) The effect of varying the closed loop gain G. dashed line: G=2; solid line: G=1; dash-dot line: G=0.5; dotted line: G=0. c) The effect of varying the level of loop inhibition β. dashed line: $\beta = 0$; solid line: $\beta \ll \Delta c$; dash-dot line: $\beta < \Delta c$; dotted line: $\beta \geq \Delta c$. See text for details.

1.4 What is positive feedback good for?

The block diagram shown in figure 6a represents a simple behavioural system. Blocks f, b and m, represent neural populations, which for the purpose of this discussion can be thought of as leaky intergrators [Arbib, 2003] (see methods section), with an output limited to a minimum firing rate of zero, and a maximum of y_{max}. A salience signal c representing the sensory and/or motivational drive for an action, is fed into a closed loop formed by blocks f and b, the output of which is passed to block m, which provides the motor signal y^m, that drives the action. Block m also receives an inhibitory signal θ (assumed constant), which acts as a threshold to ensure that no action is produced until the output of the closed loop y^f exceeds a critical value.

This architecture is loosely based on the oculomotor system (as shown in figure 1), with the single loop formed by f and b representing the combined effect of the SC-SC, SC-TH-FEF-SC, and FEF-TH-FEF loops, and θ representing the threshold effect of the omni-pause neurons in the saccadic generator circuit. Accordingly, the signal β represents the inhibitory influence of the BG on these loops, the effect of which we shall consider shortly.

We first consider the effect of the gains w^{fb}, and w^{bf}, which represent the synaptic weights of the projection from f to b and from b to f respectively. The closed loop gain G,

of the sub-system formed by f and b is given by

$$G = w^{fb} w^{bf} \tag{1}$$

Figure 6b shows the response of the system in figure 6a, to a step change in salience of Δc, for different values of G. For $G > 1$, y^f is unstable and grows exponentially before saturating at y_{max}, so that action is guaranteed provided the selection threshold θ is less than y_{max}. In this situation activity in the loop is self-sustaining, so that even when the salience signal returns to zero, the output of f remains saturated. For $G = 1$, y^f is marginally stable and increases linearly, also reaching saturation. For $G < 1$, y^f is stable and has an equivalent open-loop gain of $1/(1 - G)$, so that the final value of y^m is not guaranteed to reach saturation, but instead depends on the size of the salience signal c. Under these conditions, the output of f tracks the salience signal, returning to zero when the salience signal does so.

This simple circuit demonstrates a potential benefit that positive feedback can add to a selection system, namely the ability to raise a salience signal to the threshold for action, regardless of the size of that signal. Unchecked, this amplification will cause even the weakest of salience signals to trigger its corresponding behaviour, so that a system like this will seldom be at rest. This may upon first consideration sound rather inefficient, however, ethological models suggest such a scheme underlies animal behaviour. As Roeder [1975] points out:

> animals are usually 'doing something' during most of their waking hours, especially when in good health and under optimal conditions.

One potential benefit that arises from this tendency to act, is that problems are dealt with before they become unmanageable. For instance, in the absence of any other deficits, a mildly hungry animal will set about finding, and consuming food, thus ensuring that its hunger is sated before its energy levels become dangerously low. Accordingly, McFarland [1971] has shown that a hypothetical model of action selection incorporating positive feedback, is able to account for animal feeding patterns. By guaranteeing that motor signals reach saturation, positive feedback acts to decouple the magnitude of a response from the magnitude of the salience driving it, so that, continuing the example, an animal actively pursuing food, will do so in much the same way regardless of how hungry it actually is.

We now consider the effect of the inhibitory input β. Figure 6c shows the response of the system, to a step change in salience of Δc, with the weights $w^{fb} = w^{bf} = 1$ (and hence $G = 1$), for different values of β. When the inhibitory input to the loop is greater or equal to the salience signal i.e., $\beta \geq \Delta c$, the positive feedback is effectively disabled because the input to b is zero or less. Consequently, the system behaves like a first order system, with its output settling at the level of its input. Under these circumstances, action is not guaranteed and will depend upon the magnitude of the salience signal c. For $\beta < \Delta c$ the feedback becomes active as soon as y^f exceeds β, causing a linear increase in y^f with a rate determined by the difference $\Delta c - \beta$, thus guaranteeing that y^m reaches y_{max}, and overcomes the selection threshold. The inhibitory input also provides a means of overcoming the self-sustaining property of the loop, causing activity to decays linearly at a rate, again determined by $\Delta c - \beta$, when the salience signal returns to zero. From this it is clear that β acts as both a threshold for activation of the PFBL, and a rate controller for the evolution of activity in the loop.

Having explored the properties of a single PFBL under inhibitory control, we now present our oculomotor model (as pictured in figure 1), which in essence has a system of loops like those in figure 6, each one corresponding to a different spatial coordinate. A key difference is that for the oculomotor model, each loop's β input is determined by activity in that and all other loops, and also by the level of simulated DA.

2 Methods

Space limits preclude a full description of the model we developed, so we instead direct the reader to the papers from which the various sub-models were derived, and highlight any modifications made to those models by us.

The FEF and TH, were both modelled as a 20×20 element array of leaky integrators [Arbib, 2003], each of which was governed by the following equation:

$$\tau \dot{a} = u - a \tag{2}$$

where a represents cell activation, u the total post-synaptic current generated by afferent input to the cell, and τ represents a decay constant that depends on cell membrane properties. A piecewise linear output function was used, so that a neuron's output y, is proportional to its activation a, and has a maximum and minimum firing rate of y_{max} and y_{min} respectively.

The BG, SC and SG models of Gurney, Arai and Gancarz, each use a variation on this neural representation, with the main differences being the inclusion of reversal potentials, and the use of different output functions (e.g., sigmoidal).

The SC's layers were modelled as 20×20 element arrays as described by Arai et al.but the logarithmic mapping of visual space they used, was abandoned in favour of a simpler linear mapping. Consequently it was possible to tune the intrinsic SC weights by hand, avoiding the use of the training scheme implemented by Arai. Despite this, the motor layer weights followed the same general pattern as those used in the Arai model, namely the Mexican-hat profile, with short range excitatory and long range inhibitory connections. In addition to the visual and motor layers specified by Arai et al., we added an extra layer intended to reproduce only a motor burst at the time of saccade initiation, as opposed to the full VM activity Arai's model is intended to recreate. We refer to the layer specified by Arai and ourselves as the build-up and burst layers respectively, these being terms readily used to describe activity seen in the SC motor layer [Munoz and Wurtz, 1995]. The connectivity of the burst layer was identical to that of the build-up layer, except that rather than receiving shunting inhibition from the BG, it receives additive inhibition.

The SG model was recreated exactly as specified by Gancraz, and the output of this model was used to drive a lumped model of the oculomotor plant, which was represented as a

second order dynamic system. The SG circuit contains two separate sub-systems for the control of horizontal and vertical movements. The FEF, and the burst and motor layers of the SC, send excitatory projections to both of these, with the weights from a given element being proportional to its horizontal and vertical position in the 20×20 array.

The BG model was implemented as a 20×20 element array, and so had 400 channels as opposed to the 6 channel model used by Gurney et al. This necessitated a change in the STN, GPe and SNr layers of the model, which had to be more coarsely coded than the striatal layers (consistent with anatomy; [Oorschot, 1996]) in order that a winning channel be able to significantly influence activity in loosing channels. Consequently the projections from the 20×20 element D1, D2, FEF and TH layers had to be mapped onto the 10×10 layers that we used for the STN, GPe and SNr. Similarly projections from the 10×10 element SNr layer had to be mapped onto the 20×20 element TH and SC layers. We therefore devised a scheme for specifying weights between layers of different dimensions. This consisted of first normalising the coordinates of the array elements in the source and target layers (as specified by row and column indices). To calculate the weight between an element in the source layer and one in target layer, we calculated the Euclidean distance between their normalised coordinates, and entered it into the following formula to calculate the weight w, between the two cells:

$$w = ke^{-\frac{d^2}{2\sigma}} \tag{3}$$

where d is the normalised distance between source and target elements, and σ is a constant determining receptive field size. This gives rise to a gaussian mapping between the two layers, with cells occupying the same relative positions in their respective layers having a connecting weight equal to k, and with weights dropping off to 0 as the relative separation of cells increases.

We simulated the VDU display used by Briand et al. [1999] by generating a 20×20 input array, with the fixation and target stimuli represented by values of 0.5, and all other locations represented by 0. This array provided tonic input to the FEF representing the Y cell retinal signal relayed by posterior cortices. A simplified model of retinal processing was used to reproduce the X cell phasic signal that the SC receives. This was constructed from two 20×20 array elements of leaky integrators, each of which received input from the VDU simulation. The two layers had different time constants (1 and 5 ms), and the slower inhibited the faster, so that following a step increase in input, the output of the fast layer increased briefly, before being suppressed by the slower layer.

The whole model was solved in discrete-time with a timestep of 1ms, and using a zero-order-hold approximation. The weights connecting the various components of the model were tuned by hand to reproduce activity patterns consistent with those recorded from healthy primates performing a reflexive saccade. Following Gurney et al. [2001] we then produced a PD-like state by reducing the level of simulated DA.

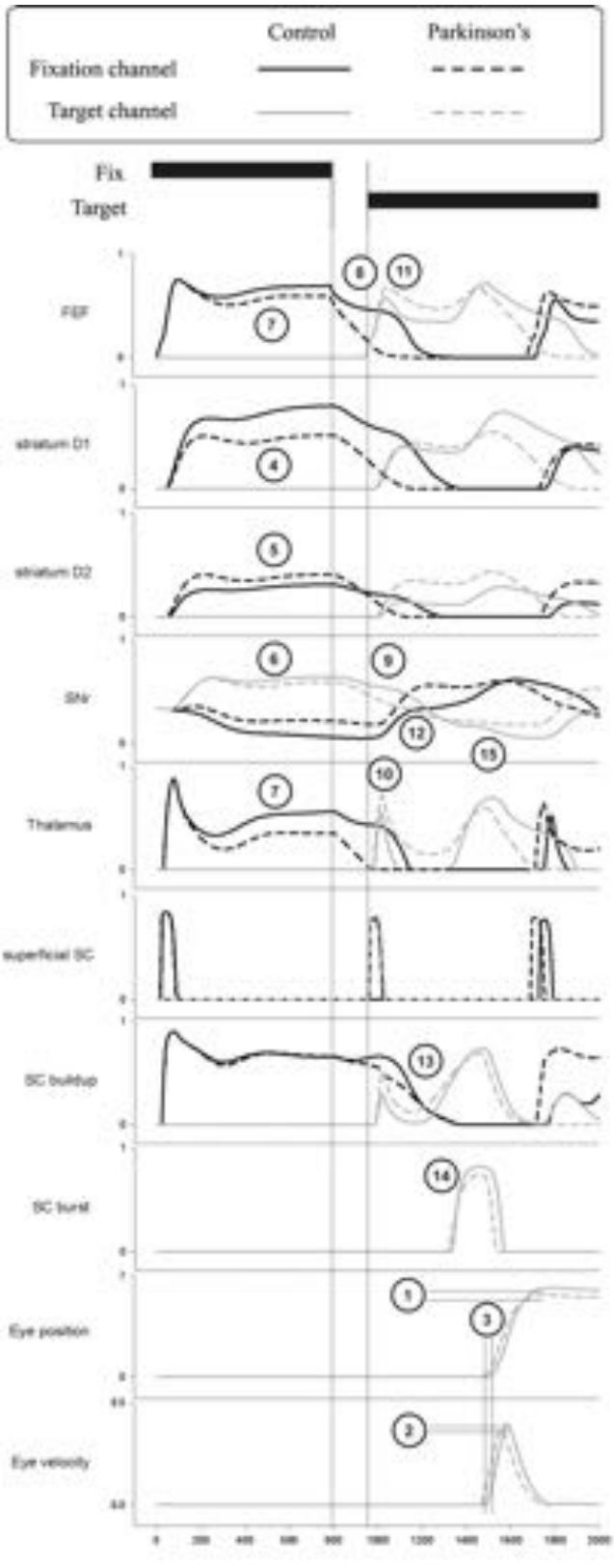


Figure 7: Results from two simulation runs. See text for details.

3 Results

Model activity is shown in figure 7, for two simulation runs, one with normal dopamine levels, and one with 1/4 of that value, these representing controls and PD patients respectively. The plots show the output of several (but not all) model layers. For those layers with a grid representation, the traces given correspond to the activity in the cells aligned with the fixation and target stimulus prior to saccade generation. Numbers in parentheses refer to points of interest marked on the plots.

3.1 Normal operation

For the control case, the onset of the fixation stimulus causes phasic activation of the retina which enters the system of PFBLs via the superficial SC. At the same time, tonic visual activity enters the system via the FEF. This combination provides salience input to the foveal channel of the BG, and with no competing activity, the BG select this as the winning channel as indicated by the reduction, and increase in SNr activity in the fixation and target channels respectively. Although the system of loops in the oculomotor model are not strictly equivalent to the simple system described in figure 6a (largely due to cross-channel communication and the use of shunting inhibition), the system behaves in a similar way to a single PFBL with a gain of less than 1. Consequently, activity in FEF, TH, and the SC build-up layer settles on a value below saturation. When the fixation stimulus is extinguished this activity begins to decay (8).

The onset of the target produces the same phasic and tonic drive to the target channel of the system. However because the BG still have the foveal channel selected, SNr input to TH and SC is elevated causing TH and SC build-up activity to return to 0 and near-zero activation after the phasic retinal input has decayed. By silencing the TH layer, BG output disables both the SC-TH-FEF-SC and FEF-TH-FEF loops, and significantly reduces activity in the SC-SC loop. Despite this, the FEF continues to provide tonic drive to the target channel of the BG, causing BG disinhibition to eventually switch to the target channel. This reduction in inhibition allows activity in the SC build-up layer to rise again, causing reactivation of TH, and hence the reactivation of positive feedback between layers. The system enters a brief period during which it behaves like a single PFBL of gain greater than 1. Target channel activity continues to increase in FEF, TH and SC build-up until, it reaches reaches a sufficient level to overcome the additive SNr inhibition to the SC burst layer. This provides a boost to FEF and SC build-up drive to the SG, that is sufficient to overcome OPN activity, and hence trigger a saccade, whereupon activity in the target channel begins to decay, on account of negative feedback from the SG to the motor layers of SC, and because visual drive to the target channel is lost as the eye begins to move.

3.2 Comparison of normal and pathological operation

In the PD-like case, activity proceeds in much the same way with the following differences. Low DA causes abnormal D1 and D2 activations (4 & 5) that result in extra inhibition on the selected fixation channel (6). This acts to reduce steady-state fixation activity in the in FEF and TH (7). Loop activity still persists when the fixation stimulus is extinguished, but its decay rate is higher than in controls (8) on account of the extra inhibition in the channel. Because fixation activity is lower in PD just prior to target onset, they have less inhibition in the target channel than controls do (9). Consequently, target activity in TH is not reduced to zero as in controls, meaning that positive feedback persists. This leads to a stronger target response in PD (10 & 11), which causes faster selection in the BG (as indicated by SNr channels crossing (12)). Consequently, the build-up of the motor burst occurs earlier in PD, triggering a saccade with a shorter latency than controls (3). However PDs incomplete disinhibition (15) prevents a burst of standard amplitude (14). The magnitude of the vector sent to the SG is therefore less than normal, and as the Gancarz model is sensitive to the size of the signal driving it, the resulting saccade has a reduced velocity (2), and is hypometric (1).

4 Discussion

We have shown how the BG may resolve the competition that takes place between a fixated stimulus and a suddenly appearing peripheral stimulus, and shown how abnormal BG function can affect this process. Future work will seek to reproduce the results of oculomotor experiments that more directly test the notion that the BG are involved in decision making. For instance, Ratcliff et al. [2003], have shown that the growth rate of SC motor activity is inversely proportional to task difficulty, an idea that is consistent with the diffusion model [Ratcliff, 1978], a psychological model of decision making. The diffusion model has at its heart, the idea that decisions are reached by accumulating evidence in favour of a decision until some critical threshold is reached. The model assumes that the rate of evidence accumulation depends on the quality of sensory information extracted from the environment, but the fact that a human subject can adjust their speed accuracy trade-off in response to verbal commands, suggests that the rate of growth in the oculomotor system is not determined by stimulus properties alone. We have shown that the rate of growth in a PFBL is related to the difference between excitatory and inhibitory input to that loop (Figure 6). By delaying the evolution of positive feedback activity, the BG may therefore be able to provide an animal with more time to gather sensory evidence, allowing a trade-off between speed and accuracy.

The oculomotor striatum receives significant input from frontal cortices known to encode current behavioural goals, and neurons found in the striatum, display a high degree of plasticity. This and related findings, have led to the suggestion that the BG may actually embody a type of reinforcement learning controller, with good outcomes acting to increase the likelihood that preceding actions are repeated [Suri and Schultz, 1999]. This raises the interesting possibility, that learning determines the level of BG disinhibition, and thus optimises the RT of an animal to a given situation (including an infinite RT, i.e., witholding a response). Future work will therefore aim to test the possibility that the BG control

of PFBLs can be thought of as a physical instantiation of an *adaptive* diffusion model.

Acknowledgments

This work was supported by an EPSRC studentship.

References

[Arai *et al*., 1994] K. Arai, E.L. Keller, and J. Edelman. Two-dimensional neural network model of the primate saccadic system. *Neural Networks*, 7(6/7):1115–1135, 1994.

[Arbib, 2003] M. Arbib. The handbook of brain theory and neural networks. MIT Press, Cambridge, Mass., 2003.

[Basso and Wurtz, 2002] M.A. Basso and R.H. Wurtz. Neuronal activity in substantia nigra pars reticulata during target selection. *J Neurosci*, 22(5):1883–1894, 2002.

[Briand *et al*., 1999] K.A. Briand, D. Strallow, W. Hening, H. Poizner, and A.B. Sereno. Control of voluntary and reflexive saccades in parkinson's disease. *Exp Brain Res*, 129(1):38–48, 1999.

[Briand *et al*., 2001] K.A. Briand, W. Hening, H. Poizner, and A.B. Sereno. Automatic orienting of visuospatial attention in parkinson's disease. *Neuropsychologia*, 39(11):1240–1249, 2001.

[Gancarz and Grossberg, 1998] G. Gancarz and S. Grossberg. A neural model of the saccade generator in the reticular formation. *Neural Netw*, 11(7-8):1159–1174, 1998.

[Gurney *et al*., 2001] K. Gurney, T.J. Prescott, and P. Redgrave. A computational model of action selection in the basal ganglia. ii. analysis and simulation of behaviour. *Biol Cybern*, 2001.

[Haber and McFarland, 2001] S. Haber and N.R. McFarland. The place of the thalamus in frontal cortical-basal ganglia circuits. *Neuroscientist*, 7(4):315–324, 2001.

[Hanes and Schall, 1996] D. P. Hanes and J. D. Schall. Neural control of voluntary movement initiation. *Science*, 274(5286):427–430, 1996.

[Harting *et al*., 2001] J.K. Harting, B.V. Updyke, and D.P. Van Lieshout. Striatal projections from the cat visual thalamus. *Eur J Neurosci*, 14(5):893–896, 2001.

[Hikosaka *et al*., 2000] O. Hikosaka, Y. Takikawa, and R. Kawagoe. Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiol Rev*, 80(3):953–978, 2000.

[Kennard and Lueck, 1989] C. Kennard and C.J. Lueck. Oculomotor abnormalities in diseases of the basal ganglia. *Rev Neurol (Paris)*, 145(8-9):587–595, 1989.

[McFarland, 1971] D. McFarland. Feedback mechanisms in animal behavior. Academic Press, New York, 1971.

[Meredith and Ramoa, 1998] M.A. Meredith and A.S. Ramoa. Intrinsic circuitry of the superior colliculus: pharmacophysiological identification of horizontally oriented inhibitory interneurons. *J Neurophysiol*, 79(3):1597–1602, 1998.

[Munoz and Istvan, 1998] D.P. Munoz and P.J. Istvan. Lateral inhibitory interactions in the intermediate layers of the monkey superior colliculus. *J Neurophysiol*, 79(3):1193–1209, 1998.

[Munoz and Wurtz, 1995] D.P. Munoz and R.H. Wurtz. Saccade-related activity in monkey superior colliculus. i. characteristics of burst and buildup cells. *J Neurophysiol*, 73(6):2313–2333, 1995.

[Oorschot, 1996] D. E. Oorschot. Total number of neurons in the neostriatal, pallidal, subthalamic, and substantia nigral nuclei of the rat basal ganglia: a stereological study using the cavalieri and optical disector methods. *J Comp Neurol*, 366(4):580–599, 1996.

[Ratcliff *et al*., 2003] R. Ratcliff, A. Cherian, and M. Segraves. A comparison of macaque behavior and superior colliculus neuronal activity to predictions from models of two-choice decisions. *J Neurophysiol*, 90(3):1392–1407, 2003.

[Ratcliff, 1978] R. Ratcliff. A theory of memory retrival. *Psychological Reviews*, 85:59–108, 1978.

[Redgrave *et al*., 1999] P. Redgrave, T.J. Prescott, and K. Gurney. The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89(4):1009–1023, 1999.

[Roeder, 1975] K. Roeder. Feedback, spontaneous activity, and behaviour. In G. Baerends, C. Beer, and A. Manning, editors, *Function and Evolution in Behaviour. Essays in Honour of Professor Niko Tinbergen, F.R.S*. Clarendon Press, Oxford, 1975.

[Schiller and Malpeli, 1977] P.H. Schiller and J.G. Malpeli. Properties and tectal projections of monkey retinal ganglion cells. *J Neurophysiol*, 40(2):428–445, 1977.

[Sommer and Wurtz, 2004] M.A. Sommer and R.H. Wurtz. What the brain stem tells the frontal cortex. i. oculomotor signals sent from superior colliculus to frontal eye field via mediodorsal thalamus. *J Neurophysiol*, 91(3):1381–1402, 2004.

[Sparks, 2002] D.L. Sparks. The brainstem control of saccadic eye movements. *Nat Rev Neurosci*, 3(12):952–964, 2002.

[Stein, 1993] B.E. Stein. *The merging of the senses / Barry E. Stein and M. Alex Meredith*. Cognitive neuroscienceseries. MIT Press, 1993.

[Suri and Schultz, 1999] R.E. Suri and W. Schultz. A neural network model with dopamine-like reinforcement signal that learns a spatial delayed response task. *Neuroscience*, 91(3):871–890, 1999.

[Windhorst, 1996] U. Windhorst. On the role of recurrent inhibitory feedback in motor control. *Prog Neurobiol*, 49(6):517–587, 1996.

When and When Not to Use Your Subthalamic Nucleus: Lessons From A Computational Model of The Basal Ganglia

Michael J. Frank
Department of Psychology and Center for Neuroscience
University of Colorado at Boulder
345 UCB, Boulder, CO 80309
frankmj@psych.colorado.edu

Abstract

The basal ganglia (BG) coordinate response selection processes by facilitating adaptive frontal motor commands while suppressing others. In previous work, a neural network model of the BG accounted for response selection deficits associated with BG dopamine depletion in Parkinson's disease. Novel predictions from this model have been subsequently confirmed in Parkinson patients and in healthy participants taking low doses of dopamine medications. Nevertheless, one clear limitation of the model is in its omission of the subthalamic nucleus (STN), a key BG structure that participates in both motor and cognitive processes. Here I include the STN and show that by modulating *when* a response is executed, it reduces premature responding and therefore has substantial effects on *which* response is ultimately selected, particularly when there are multiple competing responses. The model accurately captures the dynamics of activity in various BG areas during response selection. Simulated dopamine depletion results in emergent oscillatory activity in BG structures, which has been linked with Parkinson's tremor. Finally, the model accounts for the beneficial effects of STN lesions on these oscillations, but suggests that this benefit may come at the expense of impaired decision making.

1 Introduction

How the brain supports response selection, or decision making, is a challenge for both artificial intelligence and neuroscience communities. Based on a wealth of data, the basal ganglia (BG) are thought to play a principle role in these processes. The BG are closely interconnected with motor cortex, and are thought to modulate the execution of motor commands. Interestingly, circuits linking the BG with more cognitive areas of frontal cortex (e.g., prefrontal) are strikingly similar to those observed in the motor domain [Alexander *et al.*, 1986], raising the possibility that the BG participate in cognitive decision making in an analogous fashion to their role in motor control [Middleton and Strick, 2002].

The standard model proposes that two pathways in the BG system independently act to selectively facilitate the execution of the most appropriate cortical motor command, while suppressing competing commands [Albin *et al.*, 1989; Mink, 1996]. The general aspects of this model have been successfully leveraged to explain various motor deficits observed in patients with BG dysfunction. Recently, several researchers have pointed out that the simple aspects of this model are inadequate, and that a more advanced dynamic conceptualization of BG function is required [Gurney *et al.*, 2001; Bar-Gad *et al.*, 2003]. In particular, one key question is that if the BG participate in response selection, then how do they *learn* which response has the highest value?

1.1 A Model of Reinforcement Learning and Decision Making in PD

Previous computational modeling of the basal ganglia / dopamine system provided an explicit formulation that ties together various cognitive deficits in Parkinson's disease (PD) [Frank, 2005]. This model simulated the effects of tonic and phasic effects of dopamine on systems-level activity in the direct and indirect pathways of the basal ganglia, which are thought to independently facilitate or suppress cortical motor commands. More specifically, two main projection pathways from the striatum go through different BG output structures on the way to thalamus and up to cortex (Figure 1a). Activity in the direct pathway sends a "Go" signal to facilitate the execution of a response considered in cortex, whereas activity in the indirect pathway sends a "NoGo" signal to suppress competing responses. Dopamine modulates the relative balance of these pathways by exciting Go cells while inhibiting NoGo cells. This effect is dynamic, such that transient increases in DA leads to more Go and less NoGo, and vice versa for decreases [Frank, 2005].

The aim of the computational simulations was to explore the role of these BG dynamics in cognitive reinforcement learning, and how this is impaired in PD [Frank, 2005]. Specifically, the model (figure 1b) addressed how phasic changes in DA during error feedback are critical for modulating Go/NoGo representations in the BG that facilitate or suppress the execution of motor commands. The main assumption is that during positive and negative feedback (e.g., correct or incorrect), bursts and dips of DA occur that drive

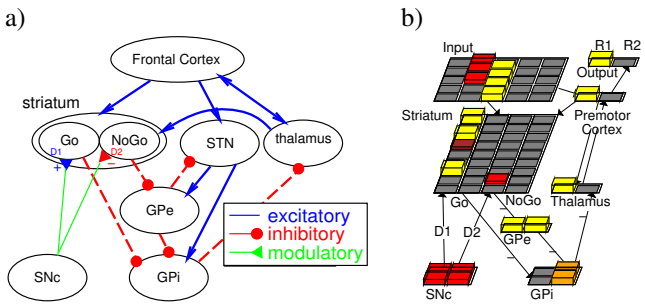


Figure 1: **a)** The striato-cortical loops, including the direct ("Go") and indirect ("NoGo") pathways of the basal ganglia. The Go cells disinhibit the thalamus via GPi, thereby facilitating the execution of an action represented in cortex. The NoGo cells have an opposing effect by increasing inhibition of the thalamus, suppressing actions from getting executed. Dopamine from the SNc projects to the dorsal striatum, causing excitation of Go cells via D1 receptors, and inhibition of NoGo via D2 receptors. GPi: internal segment of globus pallidus; GPe: external segment of globus pallidus; SNc: substantia nigra pars compacta; STN: subthalamic nucleus. **b)** The Frank (2005) neural network model of this circuit (squares represent units, with height reflecting neural activity). The Premotor Cortex selects an Output response via direct projections from the sensory Input, and is modulated by the BG projections from Thalamus. Go units are in the left half of the Striatum layer; NoGo in the right half, with separate columns for the two responses (R1 and R2). In the case shown, striatum Go is stronger than NoGo for R1, inhibiting GPi, disinhibiting Thalamus, and facilitating R1 execution in cortex. A tonic level of dopamine is shown in SNc; a burst or dip ensues in a subsequent error feedback phase (not shown), driving Go/NoGo learning. The contributions of the STN were omitted from this model, but are explored later.

learning about the response. This assumption was motivated by a large amount of evidence for bursts and dips of DA during rewards or their absence in monkeys [Schultz, 2002] which have also been inferred to occur in humans for positive and negative feedback [Holroyd and Coles, 2002; Frank *et al.*, in press]. These phasic changes in DA modulate neuronal excitability, and may therefore act to reinforce the efficacy of recently active synapses, leading to the learning of rewarding behaviors. In the model, "correct" responses are followed by transient increases in simulated DA that enhance synaptically driven activity in the direct/Go pathway, while concurrently suppressing the indirect/NoGo pathway. This drives Go learning, and enables the model to facilitate responses that on average result in positive feedback. Conversely, after incorrect responses phasic dips in DA release the NoGo pathway from suppression, increasing its activity and driving NoGo learning. Without ever having access to a supervised training signal as to which response *should* have been selected, over the course of training intact networks nevertheless learned how to respond in complex probabilistic classification tasks, similarly to healthy participants. When 75% of units in the SNc DA layer of the model were lesioned to simulate the approximate amount of damage in PD patients, the model was impaired similarly to patients.

The details of the BG model are described in Frank (2005). In brief, the premotor cortex represents and "considers" two possible responses (R1 and R2) for each input stimulus. The BG system modulates which one of these responses is facilitated and which is suppressed by signaling Go or NoGo to each of the responses. The four columns of units in the striatum represent, from left to right, Go-R1, Go-R2, NoGo-R1 and NoGo-R2. Go and NoGo representations for each response compete at the level of GPi, such that stronger Go representations lead to disinhibition of the corresponding column of the thalamus, which in turn amplifies and facilitates the execution of that response in premotor cortex. Concurrently, the alternative response is suppressed.

Striatal Go/NoGo representations are learned via phasic changes in simulated dopamine firing in the SNc layer during positive and negative reinforcement. After correct responses, increases in DA firing excite Go units for the just-selected response, while suppressing NoGo units, via simulated D1 and D2 receptors. Conversely, decreases in DA after incorrect responses results in increased NoGo activity for that response. This DA modulation of Go/NoGo activity drives learning as described above.

1.2 Modeling Dopaminergic Medication Effects on Cognitive Function in PD

The same model was used to explain certain negative effects of dopaminergic medication on cognition in PD [Frank, 2005]. While medication improves cognitive performance in some tasks, it actually tends to impair performance in probabilistic reversal learning [Cools *et al.*, 2001]. In order to simulate medication effects, it was hypothesized that medication increases the tonic level of DA, but that this interferes with the natural biological system's ability to dynamically regulate phasic DA changes. Specifically, phasic DA dips during negative feedback may be partially blocked by DA agonists that continue to bind to receptors. When this was simulated in the model, selective deficits were observed during probabilistic reversal, despite equivalent performance in the acquisition phase [Frank, 2005], mirroring the results found in medicated patients. Because increased tonic levels of DA suppressed the indirect/NoGo pathway, networks were unable to learn "NoGo" to override the prepotent response learned in the acquisition stage. This account is consistent with similar reversal deficits observed in healthy participants administered an acute dose of bromocriptine, a D2 agonist [Mehta *et al.*, 2000].

1.3 Empirical Tests of the Model

Recently, we have tested various aspects of the hypothesized roles of the basal ganglia / dopamine system across both multiple cognitive processes. First, we demonstrated support for a central prediction of our model regarding dopamine involvement in "Go" and "NoGo" cognitive reinforcement learning [Frank *et al.*, 2004; Frank, 2005]. We tested Parkinson patients on and off medication, along with healthy senior control participants. We predicted that decreased levels of dopamine in Parkinson's disease would lead to spared NoGo learning, but impaired Go learning (which depends on DA bursts). We further predicted that dopaminergic medication should alleviate the Go learning deficit, but would block the effects of dopamine dips needed to support NoGo learning.

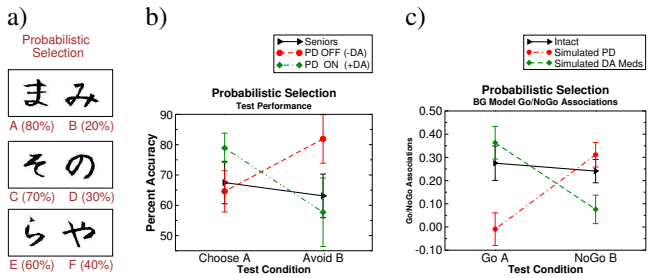


Figure 2: **a)** Example stimulus pairs (Hiragana characters) used in the cognitive probabilistic learning task, designed to minimize verbal encoding. One pair is presented per trial, and the participant makes a forced choice. The frequency of positive feedback for each choice is shown. **b)** Novel test pair performance in Parkinson patients on and off medication (Frank, Seeberger & O'Reilly, 2004), where choosing A depends on having learned from positive feedback, while avoiding B depends on having learned from negative feedback. **c)** This pattern of results was predicted by the Frank (2005) model. The figure shows Go - NoGo associations for stimulus A, and NoGo - Go associations for stimulus B, recorded from the model's striatum after having been trained on the same task used with patients. Error bars reflect standard error across 25 runs of the model with random initial weights.

Results were consistent with these predictions (Figure 2). In a probabilistic learning task, all patients and aged-matched controls learned to make choices that were more likely to result in positive rather than negative reinforcement. The difference was in their strategy: patients taking their regular dose of dopaminergic medication implicitly learned more about the positive outcomes of their decisions (i.e., they were better at Go learning), whereas those who had abstained from taking medication implicitly learned to avoid negative outcomes (better NoGo learning). Age-matched controls did not differ in their tendency to learn more from the positive/negative outcomes of their decisions. We have also found the same pattern in young healthy participants administered dopamine D2 receptors agonists and antagonists [Frank and O'Reilly, submitted]. Again, dopamine increases improved Go learning and impaired NoGo learning, while decreases had the opposite effect. Further, the same effects extended to a higher level attentional task that required paying attention to task-relevant (i.e., positively valenced) information while ignoring distracting (negative) information. Finally, this model accurately predicted the pattern of event-related potentials recorded from healthy participants that were biased to learn more from either positive or negative reinforcement [Frank *et al.*, in press].

2 Integrating Contributions of the Subthalamic Nucleus in the Model

Despite its success in capturing dopamine-driven individual differences in learning and attentional processes, the above model falls short in its ability to provide insight into BG dynamics that depend on the subthalamic nucleus (STN). The model was designed to simulate how the BG can learn to selectively facilitate (Go) one response while selectively suppressing (NoGo) another. Because the projections from the STN to BG nuclei (GPe and GPi) are diffuse [Mink, 1996; Parent and Hazrati, 1995], it may not be well suited to provide selective (focused) modulation of specific responses, and was therefore omitted from the model. Instead the model simulated the focused projections from striatum to GPi and GPe, as well as the focused projections from GPe to GPi, to demonstrate how direct and indirect pathways may compete with one another at the level of each response, but may act in parallel to facilitate and suppress alternative responses (see Frank (2005) for details and discussion).

Nevertheless, there is substantial evidence that the STN is critically involved in both motor control and cognitive processes [Bergman *et al.*, 1994; Boraud *et al.*, 2002; Baunez *et al.*, 2001; Karachi *et al.*, 2004; Witt *et al.*, 2004]. Further, other computational models of action selection also implicate a key role of the STN [Gurney *et al.*, 2001; Rubchinsky *et al.*, 2003; Brown *et al.*, 2004]. The present model explored the contributions of the STN within the computational framework of the previous model of cognitive reinforcement learning and decision making [Frank, 2005]. By virtue of its diffuse connectivity to BG nuclei, I argue that the STN may support more of a global modulatory signal on facilitation and suppression of *all* responses, rather than modulating the execution of any *particular* response. The simulations described below reveal that this global modulatory signal could not be replaced by a simple response threshold parameter, because its effects are dynamic as response selection processes evolve, and its efficacy depends on excitatory input from motor cortex. Further, simulated dopamine depletion in the augmented model results in emergent oscillations in the STN and BG output structures, which have been documented empirically and are thought to be associated the source of Parkinson's tremor. Finally, I show that the STN may be critical for action selection processes to prevent premature responding, so that all potential responses are considered before facilitating the most appropriate one.

The STN was included in the model in accordance with known constraints on its connectivity in BG circuitry, as depicted in Figure 3. First, the STN forms part of the "hyperdirect" pathway, so-named because cortical activity targets the STN, which directly excites GPi, bypassing the striatum altogether [Nambu *et al.*, 2000]. Thus initial activation of the STN by cortex leads to an initial excitatory drive on the already tonically active GPi, effectively making the latter structure more inhibitory on the thalamus, and therefore less likely to facilitate a response. Further, the STN gets increasingly excited with increasing cortical activity. Thus, if several competing responses are activated, the STN sends a stronger "Global NoGo" signal which allows the BG system to fully consider all possible options before sending a Go signal to facilitate the most adaptive one.

Second, the STN and GPe are reciprocally connected in a negative feedback loop, with the STN exciting the GPe and the GPe inhibiting the STN [Parent and Hazrati, 1995]. As noted above, the connections from STN to GPe are diffuse, and therefore are not likely to be involved in suppressing a specific response. Of the STN neurons that project to GPe, the vast majority also project to GPi [Sato *et al.*, 2000]. In the model, each STN neuron receives projections from two randomly selected GPe neurons. This was motivated by data

showing that multiple GPe neurons converge on a single STN neuron [Karachi *et al.*, 2004]. In contrast, each GPe neuron receives from a single randomly selected STN neuron.

2.1 BG Firing Patterns During Response Selection

The firing patterns of simulated BG structures during response selection are shown in Figure 3b. Upon presentation of a stimulus input, competing responses are simultaneously but weakly activated in motor cortex. Concurrently, response-specific NoGo signals from striatum cause GPe activity to decrease. The combined effects of initial cortical activity and decreases in GPe activity produce an initial STN surge at approximately 20 cycles of network settling. This STN activity is excitatory on GPi cells, preventing them from getting inhibited by early striatal Go signals that would otherwise facilitate response execution. However, STN activity also excites GPe neurons, which in turn inhibit the initial STN activity surge. At this point, a striatal Go signal for a particular response can then inhibit the corresponding GPi column, resulting in thalamic disinhibition and subsequent selection of that response in motor cortex. Because activity values are displayed in terms of average activity across each layer, the selection of a single motor response together with suppression of other responses results in a net decrease in average motor cortex activity. Finally, in some trials, a late striatal NoGo signal causes GPe inhibition and a second surge in STN activity.

The above description of STN dynamics is consistent with data from physiological recordings showing an early discharge in STN cells during response selection / initiation [Wichmann *et al.*, 1994], and with similar patterns evoked by cortical activity [Magill *et al.*, 2004]. Moreover, this model is an explicit implementation of existing theoretical constructs regarding the role of the STN in initial response suppression [Maurice *et al.*, 1998; Nambu *et al.*, 2002], followed by a direct pathway response facilitation, and then finally an indirect pathway response termination. An obvious question is whether this model also accounts for patterns of activity in the dopamine-depleted state, for which there is abundant data.

2.2 Dopamine Depletion is Associated with Subthalamic and Pallidal Oscillations

Dopaminergic depletion in Parkinson's disease is associated with changes in the firing patterns and activity levels in various BG nuclei [Mink, 1996; Boraud *et al.*, 2002]. Lowered dopamine levels result in excessive striatal NoGo (indirect pathway) activity as described earlier, which causes concomitant decreases in GPe and increases in GPi activity [Boraud *et al.*, 2002]. Parkinsonism is also associated with increased STN activity, thought to arise from reduced GABAergic GPe input [Miller and Delong, 1987; Delong, 1990]. DA depletion has also been reliably associated with low-rate oscillatory bursting activity in both STN and GPe, which is correlated with the development of Parkinson's tremor [Bergman *et al.*, 1994; 1998; Levy *et al.*, 2000; Raz *et al.*, 2000]. Finally, experimental STN lesions have been shown to eliminate GPe oscillations [Ni *et al.*, 2000] and reverse PD symptoms [Bergman *et al.*, 1990].

Interestingly, when Parkinson's disease was simulated in

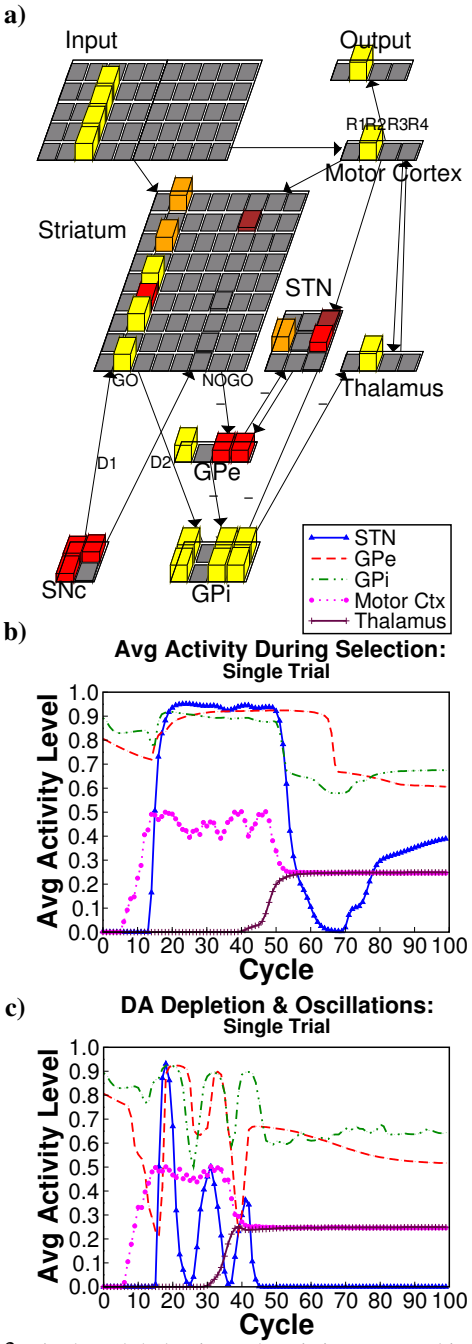


Figure 3: **a)** The subthalamic nucleus is incorporated into a scaled-up model that includes four competing responses (R1-R4). The STN receives excitatory projections from motor cortex in the "hyperdirect pathway" and excites both GPi and GPe; GPe provides inhibitory feedback on STN activity. **b)** Average layer activity levels in a single trial across network settling cycles. Initially, multiple simultaneously active motor cortex responses excite STN, which sends a "Global NoGo" signal and prevents premature responding. Sustained GPe activity subsequently inhibits STN, turning off this Global NoGo signal and allowing striatal Go signals to facilitate a response. Decreases in overall motor cortex activity are due to inhibition of the three alternative responses. Finally, striatal NoGo signals inhibit GPe, causing a second STN surge, as is observed physiologically, and is thought to terminate the executed response. **c)** Dopamine depletion leads to emergent network oscillations in STN, GPi and GPe, which have been associated with Parkinson's tremor.

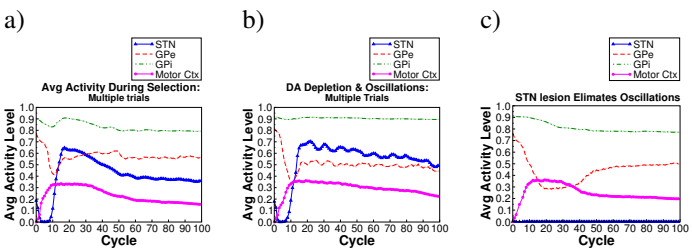


Figure 4: Average unit activity during response selection processes as a function of network settling cycles. Data are averaged across units within each area, and across 100 trials. **a)** Intact network. Dynamics are similar to those in Figure 3b, but transitions are less clearcut, since they occur at somewhat different latencies across multiple trials. **b)** Simulated Parkinsonism (DA depletion) led to increased overall GPi and STN activity, decreased GPe activity, and oscillations in both STN and GPe. **c)** STN lesions in DA-depleted networks eliminated the oscillations observed in GPe, and improved motor execution, as has been observed in experimental animals.

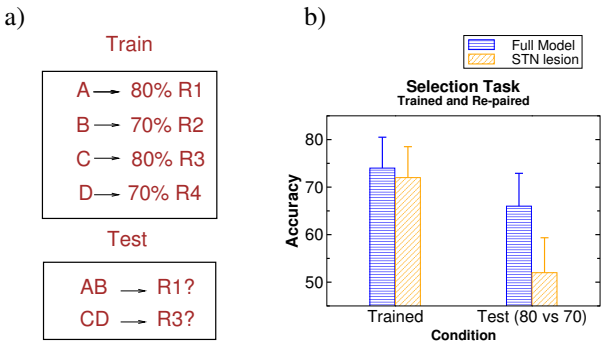


Figure 5: **a)** Response selection paradigm. Four cues are independently associated with one of four possible responses. Responses R1 and R3 are reinforced on 80% of trials in the presence of cues A and C, respectively. R2 and R4 are reinforced in 70% of trials to cues B and D. The test phase measures the network's ability to choose the 80% over the 70% response when presented with cues A and B or C and D together. **b)** Both intact networks and those with STN lesions successfully learned to choose the appropriate response for each training cue. STN lesions selectively impaired selection among two competing responses, due to premature responding before being able to integrate over all possible responses.

the model, all of these effects of simulated Parkinson's disease emerged naturally (Figure 4b). First, lesioning dopamine units in the SNc led to increased striatal NoGo activity, as described previously [Frank, 2005]. Second, this led to increased overall STN and GPi activity, and decreased GPe activity, consistent with empirical recordings. Third, and perhaps most interesting, DA depletion led to emergent network oscillations between the STN and GPe layers, which have been linked to Parkinson's tremor as described above. Similar oscillations have been previously described in a conductance model of BG function [Terman *et al.*, 2002], but these did not depend on DA depletion. Further, these oscillations dampened as the model selected a response, which is consistent with recent observations that visually guided movements suppress STN oscillations in PD [Amirnovin *et al.*, 2004], and with the fact that tremor is usually seen in the resting state. Finally, an STN "lesion" resulted in normalized GPi activity and eliminated GPE oscillations resulting from DA depletion (Figure 4c). This same pattern of results has been observed as a consequence of STN lesions in the dopamine-depleted animal [Ni *et al.*, 2000].

If STN lesions improve Parkinson symptoms, it is natural to consider what deleterious effects they might have. In other words, what is the essential computational function of the STN in action selection? Some evidence comes from the animal literature showing that STN lesions impair response selection processes, and leads to premature responding when having to suppress competing responses [Baunez *et al.*, 2001]. This leaves open the possibility that the Global NoGo signal provided by the STN is adaptive and allows the animal (or the model) sufficient time to consider all possible responses before selecting the most adaptive of them. This hypothesis is further supported by observations that STN stimulation decreases premature responding in rat [Desbonnet *et al.*, 2004]. The question is whether a formal simulation of STN involvement in BG dynamics can account for these data in a response selection paradigm.

2.3 The STN and Action Selection

To address this question, a reinforcement learning paradigm was simulated in which the network is presented with one of four cues, each represented by a column of units in the input layer. The network's task is to select one of four possible responses for each cue (Figure 5a). "Feedback" is then provided to the network by either increasing or decreasing dopamine levels. The network learns based on the difference in Go/NoGo activity levels in the response selection and feedback phase, as detailed in Frank (2005) and in the appendix.

The stimulus-response mappings are probabilistic, such that some mappings are associated with an 80% chance of dopamine bursts (and 20% dips), whereas others are associated with 70% bursts / 30% dips. All networks were trained with 20 epochs consisting of 10 trials of each stimulus cue.

To determine whether the STN is beneficial for selecting among multiple competing responses, a test phase was administered. Two cues were presented, one of which had been associated with 80% positive reinforcement for one of the responses, while the other had been associated with 70% positive reinforcement for an alternative response. Although the models had not been trained with these stimulus combinations, they should be able to select the response that was most likely to result in positive reinforcement. However, premature responding could result in selection of the 70% reinforced response, if its corresponding striatal Go signal happened to get active (due to process noise) prior to that of the 80% response. It was hypothesized that in precisely this kind of situation an initial Global NoGo signal from STN may be useful.

Simulations results were consistent with this depiction (Figure 5b). While there was no difference between networks in their ability to select the most adaptive response for each cue, models with STN lesions were impaired at choosing among two positively associated responses. This

result is consistent with the notion that the STN is critical for preventing premature responding, as networks without the STN were equally likely to choose the 70% response as the 80% response. This result is also consistent with mathematical models of optimal decision making [Bogacz *et al.*, submitted], which suggest that agents first integrate over processing noise before making a response — the present model would suggest that the STN play an important role in this speed-accuracy tradeoff. Finally, this result may explain the tendency for STN lesions to worsen, and STN stimulation to improve, premature responding in choice selection paradigms in rats [Baunez *et al.*, 2001; Desbonnet *et al.*, 2004].

3 Conclusion

How do the present simulation results provide insight into the somewhat tongue-in-cheek title of this paper – that is, when should one use or not use their subthalamic nucleus? A preliminary answer to this question may be that the STN is useful in situations that would lead to "jumping the gun" on decision making processes, by preventing premature choices. However, when excessive hesitancy is experienced, the present model would suggest turning off your STN. Future computational work may help us better understand both the therapeutic and deleterious effects of STN stimulation on motor and cognitive processes in Parkinson's disease.

A Appendix

A.1 Implementational Details

The model is implemented using the Leabra framework [O'Reilly and Munakata, 2000; O'Reilly, 2001]. Leabra uses point neurons with excitatory, inhibitory, and leak conductances contributing to an integrated membrane potential, which is then thresholded and transformed via an $x/(x+1)$ sigmoidal function to produce a rate code output communicated to other units (discrete spiking can also be used, but produces noisier results). Each layer uses a k-winners-take-all (kWTA) function that computes an inhibitory conductance that keeps roughly the *k* most active units above firing threshold and keeps the rest below threshold. Units learn based on via changes in dopamine (unsupervised), as detailed below.

The membrane potential V_m is updated as a function of ionic conductances g with reversal (driving) potentials E as follows:

$$\Delta V_m(t) = \tau \sum_c g_c(t)\overline{g_c}(E_c - V_m(t)) \tag{1}$$

with 3 channels (c) corresponding to: e excitatory input; l leak current; and i inhibitory input. Following electrophysiological convention, the overall conductance is decomposed into a time-varying component $g_c(t)$ computed as a function of the dynamic state of the network, and a constant $\overline{g_c}$ that controls the relative influence of the different conductances. The equilibrium potential can be written in a simplified form by setting the excitatory driving potential (E_e) to 1 and the leak and inhibitory driving potentials (E_l and E_i) of 0:

$$V_m^\infty = \frac{g_e\overline{g_e}}{g_e\overline{g_e} + g_l\overline{g_l} + g_i\overline{g_i}} \tag{2}$$

which shows that the neuron is computing a balance between excitation and the opposing forces of leak and inhibition. This equilibrium form of the equation can be understood in terms of a Bayesian decision making framework [O'Reilly and Munakata, 2000].

The excitatory net input/conductance $g_e(t)$ or η_j is computed as the proportion of open excitatory channels as a function of sending activations times the weight values:

$$\eta_j = g_e(t) = \langle x_i w_{ij} \rangle = \frac{1}{n}\sum_i x_i w_{ij} \tag{3}$$

The inhibitory conductance is computed via the kWTA function described in the next section, and leak is a constant.

Activation communicated to other cells (y_j) is a thresholded (Θ) sigmoidal function of the membrane potential with gain parameter γ:

$$y_j(t) = \frac{1}{\left(1 + \frac{1}{\gamma[V_m(t) - \Theta]_+}\right)} \tag{4}$$

where $[x]_+$ is a threshold function that returns 0 if $x < 0$ and x if $X > 0$. Note that if it returns 0, we assume $y_j(t) = 0$, to avoid dividing by 0. As it is, this function has a very sharp threshold, which interferes with graded learning learning mechanisms (e.g., gradient descent). To produce a less discontinuous deterministic function with a softer threshold, the function is convolved with a Gaussian noise kernel ($\mu = 0$, $\sigma = .005$), which reflects the intrinsic processing noise of biological neurons:

$$y_j^*(x) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-z^2/(2\sigma^2)} y_j(z - x)dz \tag{5}$$

where x represents the $[V_m(t) - \Theta]_+$ value, and $y_j^*(x)$ is the noise-convolved activation for that value.

A.2 Inhibition Within and Between Layers

Inhibition *between* layers (i.e for GABAergic projections between BG layers) is achieved via simple unit inhibition, where the inhibitory current g_i for the unit is determined from the net input of the sending unit.

For *within* layer inhibition, Leabra uses a kWTA (k-Winners-Take-All) function to achieve inhibitory competition among units within each layer (area). The kWTA function computes a uniform level of inhibitory current for all units in the layer, such that the $k+1$th most excited unit within a layer is generally below its firing threshold, while the kth is typically above threshold. Activation dynamics similar to those produced by the kWTA function have been shown to result from simulated inhibitory interneurons that project both feedforward and feedback inhibition [O'Reilly and Munakata, 2000]. Thus, although the kWTA function is somewhat biologically implausible in its implementation (e.g., requiring global information about activation states and using sorting mechanisms), it provides a computationally effective approximation to biologically plausible inhibitory dynamics.

kWTA is computed via a uniform level of inhibitory current for all units in the layer as follows:

$$g_i = g^\Theta_{k+1} + q(g^\Theta_k - g^\Theta_{k+1}) \tag{6}$$

where $0 < q < 1$ (.25 default used here) is a parameter for setting the inhibition between the upper bound of g_k^Θ and the lower bound of g_{k+1}^Θ. These boundary inhibition values are computed as a function of the level of inhibition necessary to keep a unit right at threshold:

$$g_i^\Theta = \frac{g_e^* \bar{g}_e (E_e - \Theta) + g_l \bar{g}_l (E_l - \Theta)}{\Theta - E_i} \quad (7)$$

where g_e^* is the excitatory net input without the bias weight contribution — this allows the bias weights to override the kWTA constraint.

In the kWTA function used here, g_k^Θ and g_{k+1}^Θ are set to the threshold inhibition value for the kth and $k+1$th most excited units, respectively. Thus, the inhibition is placed exactly to allow k units to be above threshold, and the remainder below threshold.

A.3 Learning

Synaptic connection weights were trained using a reinforcement learning version of Leabra. The learning algorithm involves two phases, and is more biologically plausible than standard error backpropagation. In the *minus phase*, the network settles into activity states based on input stimuli and its synaptic weights, ultimately "choosing" a response. In the *plus phase*, the network resettles in the same manner, with the only difference being a change in simulated dopamine: an increase of SNc unit firing from 0.5 to 1.0 for correct responses, and a decrease to zero SNc firing for incorrect responses. Connection weights are then adjusted to learn on the difference between pre and postsynaptic activation product across the minus and plus phases [O'Reilly, 1996].

References

[Albin *et al.*, 1989] R. L. Albin, A. B. Young, and J. B. Penney. The functional anatomy of basal ganglia disorders. *Trends in Neurosciences*, 12:366–375, 1989.

[Alexander *et al.*, 1986] G. E. Alexander, M. R. DeLong, and P. L. Strick. Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, 9:357–381, 1986.

[Amirnovin *et al.*, 2004] R. Amirnovin, Z. M. Williams, G. R. Cosgrove, and E. N. Eskandar. Visually guided movements suppress subthalamic oscillations in parkinson's disease patients. *Journal of Neuroscience*, 24(50):11302–11306, 2004.

[Bar-Gad *et al.*, 2003] I. Bar-Gad, G. Morris, and H. Bergman. Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Progress in Neurobiology*, 71:439–73, 2003.

[Baunez *et al.*, 2001] C. Baunez, T. Humby, D. M. Eagle, L. J. Ryan, S. B. Dunnett, and T. W. Robbins. Effects of STN lesions on simple vs choice reaction time tasks in the rat: Preserved motor readiness, but impaired response selection. *European Journal of Neuroscience*, 13:1609–16, 2001.

[Bergman *et al.*, 1990] H. Bergman, T. Wichmann, and M. R. DeLong. Reversal of experimental parkinsonism by lesions of the subthalamic nucleus. *Science*, 249:1436–8, 1990.

[Bergman *et al.*, 1994] H. Bergman, T. Wichmann, B. Karmon, and M. R. DeLong. The primate subthalamic nucleus. II. neuronal activity in the MPTP model of parkinsonism. *Journal of Neurophysiology*, 72:507–20, 1994.

[Bergman *et al.*, 1998] H. Bergman, A. Feingold, A. Nini, A. Raz, H. Slovin, M. Abeles, and E. Vaadia. Physiological aspects of information processing in the basal ganglia of normal and parkinsonian primates. *Trends in Neurosciences*, 21:32–8, 1998.

[Bogacz *et al.*, submitted] R. Bogacz, E. Brown, J. Moehlis, P. Hu, P. Holmes, and J. D. Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. submitted.

[Boraud *et al.*, 2002] T. Boraud, E. Bezard, B. Bioulac, and E. Gross. From single extracellular unit recording in experimental and human Parkinsonism to the development of a functional concept of the role played by the basal ganglia in motor control. *Progress in Neurobiology*, 66(4):265–83, 2002.

[Brown *et al.*, 2004] J. W. Brown, D. Bullock, and S. Grossberg. How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. *Neural Networks*, 17:471–510, 2004.

[Cools *et al.*, 2001] R. Cools, R. A. Barker, B. J. Sahakian, and T. W. Robbins. Enhanced or impaired cognitive function in Parkinson's disease as a function of dopaminergic medication and task demands. *Cerebral Cortex*, 11:1136–1143, 2001.

[Delong, 1990] M. R. Delong. Primate models of movement disorders of basal ganglia origin. *Trends in Neuroscience*, 13:281–5, 1990.

[Desbonnet *et al.*, 2004] L. Desbonnet, Y. Temel, V. Visser-Vandewalle, A. Blokland, V. Hornikx, and H. W. Steinbusch. Premature responding following bilateral stimulation of the rat subthalamic nucleus is amplitude and frequency dependent. *Brain Research*, 1008:198–204, 2004.

[Frank and O'Reilly, submitted] M. J. Frank and R. C. O'Reilly. A mechanistic account of striatal dopamine function in cognition: Psychopharmacological studies with cabergoline and haloperidol. submitted.

[Frank *et al.*, 2004] M. J. Frank, L. C. Seeberger, and R. C. O'Reilly. By carrot or by stick: Cognitive reinforcement learning in Parkinsonism. *Science*, 306:1940–3, 2004.

[Frank *et al.*, in press] M. J. Frank, B. S. Woroch, and T. Curran. Error-related negativity predicts reinforcement learning and conflict biases. *Neuron*, in press.

[Frank, 2005] M. J. Frank. Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and non-medicated Parkinsonism. *Journal of Cognitive Neuroscience*, 17:51–72, 2005.

[Gurney *et al.*, 2001] K. Gurney, T. J. Prescott, and P. Redgrave. A computational model of action selection in the basal ganglia. I. a new functional anatomy. *Biological Cybernetics*, 84:401–410, 2001.

[Holroyd and Coles, 2002] C. B. Holroyd and M. G. H Coles. The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109:679–709, 2002.

[Karachi *et al.*, 2004] C. Karachi, J. Yelnik, D. Tande, L. Tremblay, E. C. Hirsch, and C. Francois. The pallidosubthalamic projection: An anatomical substrate for nonmotor functions of the subthalamic nucleus in primates. *Movement Disorders*, 2004.

[Levy *et al.*, 2000] R. Levy, W. D. Hutchison, A. M. Lozano, and J. O. Dostrovsky. High-frequency synchronization of neuronal activity in the subthalamic nucleus of parkinsonian patients with limb tremor. *Journal of Neuroscience*, 20:7766, 2000.

[Magill *et al.*, 2004] Peter J. Magill, Andrew Sharott, Mark D. Bevan, Peter Brown, and J. Paul Bolam. Synchronous unit activity and local field potentials evoked in the subthalamic nucleus by cortical stimulation. *Journal of Neurophysiology*, 92(2):700–714, 2004.

[Maurice *et al.*, 1998] N. Maurice, J.-M. Deniau, J. Glowinski, and A.-M. Thierry. Relationships between the prefrontal cortex and the basal ganglia in the rat: Physiology of the corticosubthalamic circuits. *Journal of Neuroscience*, 18:9539, 1998.

[Mehta *et al.*, 2000] M. A. Mehta, R. Swainson, A. D. Ogilvie, B. J. Sahakian, and T. W. Robbins. Improved short-term spatial memory but impaired reversal learning following the dopamine D2 agonist bromocriptine in human volunteers. *Psychopharmacology*, 159:10–20, 2000.

[Middleton and Strick, 2002] F. A. Middleton and P. L. Strick. Basal-ganglia 'projections' to the prefrontal cortex of the primate. *Cerebral Cortex*, 12:926–935, 2002.

[Miller and Delong, 1987] W.C¿ Miller and M. R. Delong. Altered tonic activity of neurons in the globus pallidus and subthalamic nucleus in the primate MPTP model of parkinsonism. In M. B. Carpenter and A. Jayaraman, editors, *The Basal Ganglia*, volume II, pages 415–427. Plenum Press, New York, 1987.

[Mink, 1996] J. W. Mink. The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, 50:381–425, 1996.

[Nambu *et al.*, 2000] A. Nambu, H. Tokuno, I. Hamada, H. Kita, M. Imanishi, T. Akazawa, Y. Ikeuchi, and N. Hasegawa. Excitatory cortical inputs to pallidal neurons via the subthalamic nucleus in the monkey. *Journal of Neurophysiology*, 84:289–300, 2000.

[Nambu *et al.*, 2002] A. Nambu, H. Tokuno, and M. Takada. Functional significance of the cortico-subthalamo-pallidal 'hyperdirect' pathway. *Neuroscience Research*, 43:111–7, 2002.

[Ni *et al.*, 2000] Z. Ni, R. Bouali-Benazzouz, D. Gao, A. Benabid, and A. Benazzouz. Changes in the firing pattern of globus pallidus neurons after the degeneration of nigrostriatal pathway are mediated by the subthalamic nucleus in rat. *European Journal of Neuroscience*, 12:4338–44, 2000.

[O'Reilly and Munakata, 2000] R. C. O'Reilly and Y. Munakata. *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. MIT Press, Cambridge, MA, 2000.

[O'Reilly, 1996] R. C. O'Reilly. Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, 8(5):895–938, 1996.

[O'Reilly, 2001] R. C. O'Reilly. Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation*, 13:1199–1242, 2001.

[Parent and Hazrati, 1995] A. Parent and L. Hazrati. Functional anatomy of the basal ganglia. II. the place of subthalamic nucleus and external pallidum in basal ganglia circuitry. *Brain Research Reviews*, 20:128–54, 1995.

[Raz *et al.*, 2000] A. Raz, E. Vaadia, and H. Bergman. Firing patterns and correlations of spontaneous discharge of pallidal neurons in the normal and the tremulous 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine vervet model of parkinsonism. *Journal of Neuroscience*, 20:8559–71, 2000.

[Rubchinsky *et al.*, 2003] L. L. Rubchinsky, N. Kopell, and K. A. Sigvardt. Modeling facilitation and inhibition of competing motor programs in basal ganglia subthalamic nucleus-pallidal circuits. *Proceedings of the National Academy of Sciences*, 100:14427–32, 2003.

[Sato *et al.*, 2000] F. Sato, M. Parent, M. Levesque, and A. Parent. Axonal branching pattern of neurons of the subthalamic nucleus in primates. *Journal of Comparative Neurology*, 424:142–52, 2000.

[Schultz, 2002] W. Schultz. Getting formal with dopamine and reward. *Neuron*, 36:241–263, 2002.

[Terman *et al.*, 2002] D. Terman, J. E. Rubin, A. C. Yew, and C. J. Wilson. Activity patterns in a model for the subthalamopallidal network of the basal ganglia. *Journal of Neuroscience*, 22:2963–2976, 2002.

[Wichmann *et al.*, 1994] T. Wichmann, H. Bergman, and M. R. DeLong. The primate subthalamic nucleus. I. functional properties in intact animals. *Journal of Neurophysiology*, 72:494–506, 1994.

[Witt *et al.*, 2004] K. Witt, U. Pulkowski, J. Herzog, D. Lorenz, W. Hamel, G. Deuschl, and P. Krack. Deep brain stimulation of the subthalamic nucleus improves cognitive flexibility but impairs response inhibition in Parkinson's disease. *Archives of Neurology*, 61:697–700, 2004.

Action selection in a macroscopic model of the brainstem reticular formation

Mark Humphries, Kevin Gurney and Tony Prescott
Adaptive Behaviour Research Group
University of Sheffield
Department of Psychology, Western Bank, Sheffield, S10 2TP.
m.d.humphries@sheffield.ac.uk

Abstract

The behavioral repertoire of decerebrate and neonatal animals suggests that a relatively self-contained neural substrate of action selection may exist in the brainstem. Here we develop the hypothesis that the principal component of the substrate is the medial ponto-medullary reticular formation. Our quantitative structural model of this region, which proposes a macroscopic organisation at the level of inter-connected neural clusters, is extended to incorporate sensory input. Evidence is reviewed in support of the proposal that both input and output configurations of this region follow this organisation. To investigate how this biologically-constrained model may be configured to support action selection, a computational neural-population model of the medial reticular formation is outlined, and alternate configurations are assessed in simulation. We conclude that the configuration which most effectively supports action selection is likely to be one which represents compatible sub-actions at the cluster level; thus, co-activation of a set of these clusters would lead to the co-ordinated behavioral response observed in the animal.

1 Introduction

It is a safe assumption that action selection by animals is achieved through some neural process. Recent proposals for the neural substrate of the vertebrate action selection system have focussed on the basal ganglia - a set of fore- and mid-brain nuclei whose input, output, and inter-connections seem to be consistent with a central (as opposed to distributed) resource switching device [Mink and Thach, 1993; Redgrave *et al.*, 1999; Prescott *et al.*, 1999]. Animals which lack a functioning basal ganglia are not completely impaired, though their behavioral repertoire is undeniably limited. Thus, the basal ganglia may form a critical, but not necessary, part of the action selection neural substrate.

Decerebrate animals, those from which the entire brain anterior to the superior colliculus - including the basal ganglia - has been removed, and altricial (helpless at birth) neonates, for which the basal ganglia circuitry is not complete, are capable of expressing spontaneous behaviors and co-ordinated and appropriate responses to stimuli. The chronic decerebrate rat can, for example, spontaneously locomote, orient correctly to sounds, groom, perform co-ordinated feeding actions, and discriminate food types [Berntson and Micco, 1976; Grill and Kaplan, 2002]. Such animals clearly have some form of intact system for simple action selection, which enables them to respond to stimuli with appropriate actions that are more complex than simple spinal-level reflexes, and which enables them to sequence behaviors, as in the holding, gnawing, and chewing required for eating solid food.

Of the potential candidate structures left intact in the brainstem of decerebrate animals, we propose that the medial ponto-medullary reticular formation (RF) is the most likely substrate of a generalised simple action selection mechanism. Brainstem structures outside the RF are either motor relays (the cranial nerve nuclei), sensory relays (trigeminal nucleus), or cerebellar relays. The multitude of structures within the RF participate in REM sleep control, global neuromodulation (for example, the serotonergic raphe nuclei), oculomotor control, and, again, cerebellar relays. It is the medial core of the RF which lacks a clear functional role, and which seemingly has the circuitry necessary to perform some form of action selection. We are not proposing that the medial RF subsumes the basal ganglia's action selection role, but rather that the RF is capable of performing limited action selection in the basal ganglia's absence.

In a landmark paper, Warren McCulloch proposed that the medial RF was the substrate for the selection of an organism's global behavioral state, which was set by the RF's connections with the mid- and fore-brain [Kilmer *et al.*, 1969]. The paper also described a computational model which demonstrated that the known anatomy of the medial RF could support selection-like functions. We have previously demonstrated that an altered, optimised version of this model can support action selection in a simple robotic foraging task [Humphries *et al.*, a]. However, due to its age, inevitably the model contains incorrect assumptions about, and omits important features of, the RF (in particular the predominance of posterior projections to the spine over anterior projections to the mid- and fore-brain). In accordance with our proposal, and consistent with the dominance of spinal projections, later authors have argued that the medial RF is involved solely in motor control [Siegel, 1979]. Here we demonstrate that the medial RF can support action selection-like properties in

a simple dynamic model which explicitly addresses motor function.

2 The structural model

A review of the medial RF's anatomy led us to propose a quantitative structural model which described its neuronal organisation [Humphries *et al.*, b]. We identified two main neuron types. The projection neurons extend a bifurcating axon, predominantly caudally to the spinal cord and rostrally toward the midbrain, and make excitatory contacts on their targets via extensive collateralisation along the main axon. The inter-neurons project their axon almost entirely within the RF, predominantly along the medio-lateral axis, and make inhibitory contacts with their targets. These neurons are arranged into clusters comprising a set of projection and inter-neurons, each cluster delimited by the initial collateral from the projection neurons' axons - which occurs roughly 200μm from the initial bifurcation. This proposed cluster model of medial RF structure is explained further in Figure 1.

The quantitative structural model is as follows. Every one of the N_c clusters in the model has n neurons; the total number of neurons T within the model is thus $T = N_c \times n$. Within each cluster a certain proportion ρ of neurons are deemed to be the projection neurons, the remainder are deemed to be inter-neurons. From the data reviewed in [Humphries *et al.*, b], we set bounds $0.7 \leq \rho < 0.9$. Three parameters define the stochastic connectivity between neurons. For each projection neuron, the probability of forming a connection c between itself and another cluster is $P(c)$. Data from [Grantyn *et al.*, 1987] suggests a spatially uniform model for which we assign $P(c) = 0.25$ for all clusters. (An alternative, a distance-dependent distribution typical of many neural structures, was explored in [Humphries *et al.*, b]; here we do not consider that distribution to simplify our discussions).

If a connection is made then $P(p)$ is the probability that the projection neuron forms a connection p with a given neuron in that cluster. Finally, $P(l)$ denotes the probability of connection l between an inter-neuron and any other neuron in its cluster. All models are homogeneous with respect to intra-cluster connectivity, so that the probabilities $P(l), P(p)$ are independent of particular clusters and neurons within clusters. When we construct a particular instantiation of the structural model, the above parameters are used to define directed edges in a connectivity graph, where each vertex (node) of the graph is labeled as being either a projection or an inter-neuron.

To this existing model, we must add definitions for sensory input. Two parameters are added to the structural model to define the proportion of neurons that receive sensory input: a proportion of projection neurons ρ_s and a proportion of interneurons λ_s are defined as receiving sensory afferents within each cluster - these proportions are the same for every cluster. Given the extent and morphology of their dendritic trees, it it likely that the projection neurons within a cluster will receive synaptic input from the majority of sensory afferents contacting that cluster. In addition, projection neurons which do not respond to some form of sensory stimulation are rare [Schulz *et al.*, 1983]. Thus we set $\rho_s = 1$ throughout.

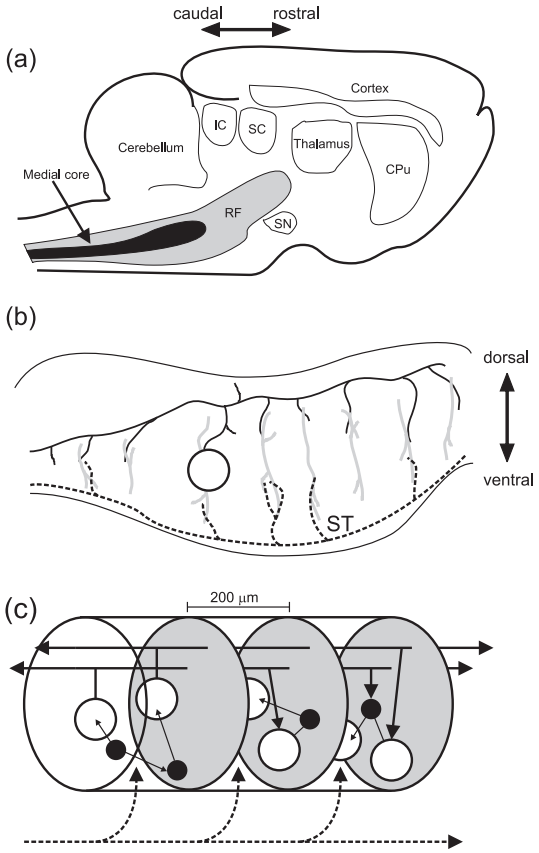


Figure 1: Schematic summary of the vertebrate reticular formation's anatomical organisation. Directional arrows apply to all panels. (a) Sagittal section of cat brain, showing relative size and location of reticular formation (RF) and medial core. Abbreviations: CPu - caudate-putamen (striatum); IC - inferior colliculus; SC - superior colliculus; SN - substantia nigra. (b) Sagittal section through the brainstem; the dendritic trees (grey lines) of the projection neurons (one cell body shown - open circle) extend throughout the medial RF along the dorso-ventral axis but extend little along the rostro-caudal axis. These dendritic trees contact axon collaterals of both ascending sensory systems (black dashed line) and far-reaching axons of the projection neurons (the axon of the depicted cell body is shown by the solid black line); ST is the spinothalamic tract. (c) The cluster model of RF organisation. The medial RF is comprised of stacked clusters (3 shown) containing medium-to-large projection neurons (open circles) and small-to-medium inter-neurons (filled circles); cluster limits (grey ovals) are defined by the initial collaterals from the projection neuron axons. Their radial dendritic fields allow sampling of ascending and descending input from both other clusters (solid black lines) and sensory systems (dashed black line). The interneurons project predominantly within their parent cluster.

Patterning of sensory inputs to the inter-neurons is unknown, but a similar argument, based on their dendritic morphology, would suggest that proportionally fewer inter-neurons than projection neurons would receive input from the same sensory afferent to their cluster. Some medium-sized cells, which could potentially be inter-neurons, do receive spinal input [Eccles *et al.*, 1976], and thus some form of sensory input to inter-neurons cannot be entirely ruled out. We must thus allow λ_s to vary over the interval $[0, 0.5]$ in a full exploration of the model. The result of these additions is that each node in the connectivity graph now has assigned to it a flag indicating the presence or absence of sensory input, which is used in the dynamic model.

3 The macroscopic dynamic model

To capture the global dynamic properties of a neural system, the use of so-called macroscopic models is often adopted [Latham *et al.*, 2000; Monteiro *et al.*, 2002]. In this approach the activity of neural populations is captured as a simplified set of ordinary differential equations (ODEs) which may reveal qualitatively similar dynamics to more complete models with individual neural elements. Here, we establish a macroscopic model of the medial RF cluster model. The macroscopic model we propose is a reduction to the level of cluster dynamics, based on the assumption that individual unit dynamics are standard leaky-integrators [Gurney *et al.*, 2001]. An additional assumption, following the structural model, is that all projection neurons are excitatory and all inter-neurons inhibitory. For cluster k, its normalised, average projection neuron output c_k is given by

$$\tau \frac{dc_k(t)}{dt} = -c_k(t) + F\Big(\bar{w}_e \sum_{j=1}^{N_c} A_{jk}\, c_j(t) + \bar{w}_i b_k i_k(t) + \rho_s u_k(t)\Big), \quad (1)$$

where τ is a time constant dictating the decay rate of the neural activity, $F(x)$ is the output function, $\bar{w}_e, \bar{w}_i$, are the mean excitatory and inhibitory weights, $c_j(t)$ is the average projection neuron output from cluster j, and $u_k(t)$ is input to the current cluster. The average inter-neuron output i_k of cluster k is given by

$$\tau \frac{di_k(t)}{dt} = -i_k(t) + F\Big(\bar{w}_e \sum_{j=1}^{N_c} C_{jk}\, c_j(t) + \bar{w}_i d_k\Big(i_k(t) - \frac{i_k(t)}{n^-}\Big) + \lambda_s u_k(t)\Big), \quad (2)$$

where $n^- = n(1-\rho)$ is the number of inter-neurons per cluster - the bracketed term containing this parameter describes the contribution of the inter-neuron population to itself. Variables A_{jk}, b_k, C_{jk}, d_k are scalars determined from the properties of the underlying structural model (section 2): A_{jk}, C_{jk} are the mean number of contacts from afferent cluster j to, respectively, the projection and inter-neurons; b_k, d_k are the mean number of contacts from inter-neurons in current cluster k to, respectively, the projection and inter-neurons in that cluster.

We use a piece-wise linear output function given by

$$F(x) = \begin{cases} 0, & \text{if } x < \epsilon; \\ m(x-\epsilon), & \text{if } \epsilon \le x \le 1/m + \epsilon \\ 1, & \text{if } x > 1/m + \epsilon \end{cases} \quad (3)$$

where m is slope, and ϵ the threshold of the output function. Throughout, we set $m = 1$ and $\epsilon = 0$.

4 The medial RF as an action selection system

We briefly discuss how the input and output configurations are constrained by the biological data, then move on to a consideration of how the combined structural and dynamic models may give rise to an action selection system.

4.1 Input configuration

Sensory inputs originate from ascending spinal systems, such as the spinothalamic tract depicted in Figure 1, and from brainstem relay nuclei, such as the dorsal cochlear nucleus which conveys auditory information. Axons of spinal origin and of some brainstem relay nuclei - for example, the sensory trigeminal nucleus - collateralise in a similar manner to the projection neurons' axons, sending branches perpendicular to the main axon trunk into the clusters, with each branch contacting only one cluster. The organisation of input from the other brainstem relay nuclei is unknown, but we assume it forms part of the main fiber bundles traversing the brainstem, and is therefore likely to collateralise in the same way.

The multi-modal sensory responses of projection neurons are evidence that multiple sensory input systems contact the same neuron [Scheibel, 1984]. Neighboring pairs of projection neurons have correlated activity in the waking animal, which is evidence for a common afferent input, but distal projection neuron pairs do not [Siegel *et al.*, 1981]. Thus, both the anatomical organisation and neural activity characteristics are consistent with each cluster having a unique pattern of multi-modal sensory input (and, conversely, are consistent with the assumption that such a macro-scale object - the cluster - exists as an organisational element in the RF).

We thus interpret the single input variable u_k to be a normalised scalar summation of all sensory input to cluster k. The majority of sensory inputs are assumed to be excitatory, as firing rate increases are generally reported following the presentation of stimuli. However, inhibitory responses have been reported following both visceral and somatic stimulation [Langhorst *et al.*, 1996], which may reflect either direct inhibitory input, or indirect inhibition via afferent drive of the inhibitory inter-neurons. Thus, we are not able to state definitively that sensory input is entirely excitatory, and must therefore consider u_k over the interval [-1,1] in a full exploration of the model - to simplify the discussions below, here we consider u_k only over the interval [0,1].

4.2 Output configuration

The projection neurons' targets in the cranial nerve nuclei and the spine are assumed to express the action selected by the medial RF system. Many projection neurons have correlated activity with multiple movements, and the activity of near-neighbor projection neurons often does not correlate

with the same movement or set of movements [Siegel and Tomaszewski, 1983]. Thus, the correlated activity between near-neighbor projection neurons in waking animals [Siegel *et al.*, 1981] would lead to the simultaneous recruitment of multiple muscle groups and movement types. We therefore propose that sufficient activation of a cluster's projection neurons would lead to a co-ordinated behavioral response.

Consistent with this proposal, micro-stimulation studies of the medial medullary RF have demonstrated both multiple movement and multiple muscle responses following the injection of short trains of low-amplitude current pulses [Drew and Rossignol, 1990]. The same micro-stimulation applied to the lateral medullary RF did not consistently result in movement, further evidence that the medial RF is the substrate of action selection in the brainstem.

4.3 Potential configurations as an action selection system

Here we explore how the anatomical organisation of the medial RF, as defined by structural model, and constrained by the input and output patterns just described, could be configured to act as the action selection system of the brainstem. As potential configurations are discussed, the properties of each are demonstrated by an example simulation of the macroscopic dynamic model. A structural model containing just $N_c = 3$ clusters, each nominally with $n = 100$ neurons, was constructed, with the parameter set: $P(l) = P(p) = 0.1$, as arbitrarily chosen neuron pairs are likely to have low connection probabilities [Schuz, 1995]; $\rho = 0.8$, as this is the middle of the range of projection neuron proportions; and $\lambda_s = 0$, so that we need only consider effects of sensory inputs to the projection neurons - however, note that increasing λ_s to its maximum value ($\lambda_s = 0.5$) did not alter the relative values of the output reported below.

The macroscopic dynamic model connection matrices $\mathbf{A}, \mathbf{b}, \mathbf{C}, \mathbf{d}$ for the particular structural model used here are given in Appendix A. We set $\bar{w}_e = 0.2$ - the values for $\bar{w}_i$ are discussed below - and $\tau = 0.005$. Each example simulation has the same continuous (i.e. $\mathbf{u}(t) = \mathbf{u}$) input pattern, $\mathbf{u} = [0.4\ 0.3\ 0.2]$. The ODE system described by equations (1) and (2) was solved numerically using the variable-step Runge-Kutta solver in MatLab (MathWorks), with initial conditions $c_k(0) = i_k(0) = 0$, and the cluster outputs recorded after equilibrium was reached.

Single-action configuration

Where all clusters are significantly connected to all other clusters - that is, the combined output targets of the projection neurons of a cluster covers a roughly equal sampling of all other clusters (as created by the spatially-uniform collateral model) - then a winner-takes all (WTA) type circuit could potentially result, as shown in Figure 2a. In such a circuit, the outputs of each cluster are taken to activate a complete action.

For this to be the case, it is self-evident that the projection neuron component of the cluster must receive greater input from its corresponding inter-neuron component than from the combined input of the inter-cluster connections, as otherwise the net effect of any sensory input would be excitatory in a symmetrical network. One possibility is that inter-

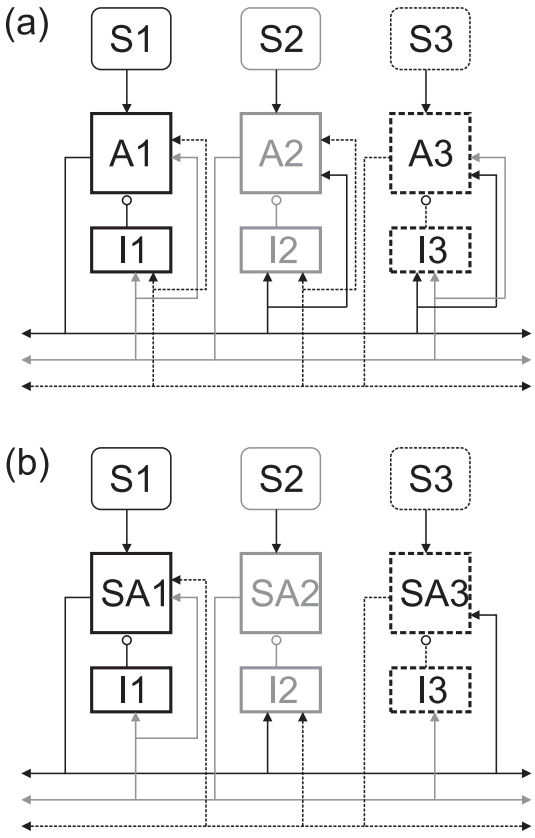


Figure 2: Potential configurations of the medial RF cluster architecture as an action selection mechanism. The different line types identify neuron populations and connections originating from the same cluster. Cluster-specific total sensory input (Sn) targets only the cluster's projection neuron component (in these examples), whose outputs drive some form of coherent behavioral response to that particular combination of sensory input. (a) A spatially-uniform distribution of inter-cluster connections may result in a winner-takes-all (WTA) type architecture, in which each cluster's projection neurons drives a complete action (An), and target other cluster's inter-neurons (In) in roughly equal proportion. Note that, to form a WTA circuit, the relative weighting of the within-cluster inter-neuron connections (inhibitory, open circles) must be greater than the projection neurons connections to other clusters (excitatory, arrows). (b) Specific wiring configurations may promote mutual excitation as well as inhibition, creating a circuit in which the sensory activation of a single sub-action (SAn) may recruit other compatible (or essential) sub-actions to be co-expressed via the inter-cluster connections between projection neurons. The combination of sub-action activations creates the coherent behavioral response observed in the animal.

cluster connections to inter-neurons have a higher weight than inter-cluster connections to projection neurons in the same target cluster. However, without detailed anatomical data on, for example, bouton counts from a single axon, there is no *a priori* reason to believe this to be true. The alternative is that the connection from the cluster's inter-neurons to its projection neurons has a relatively high weight compared to the inter-cluster connection weight, and thus afferent drive from a cluster will result in a net inhibitory effect. Synapse counts from projection neuron dendritic trees suggest that this may be the case: roughly 45% of the synapses on a projection neuron are GABAergic [Jones *et al.*, 1991] - and, therefore, inhibitory - and inter-neurons are the primary (perhaps only) source of GABAergic input, yet the proportion of inter-neurons to projection neurons is much smaller than this value. Thus, an inter-neuron input to a projection neuron would have a disproportionately larger effect than a given projection neuron input, as it forms more synapses. Therefore, we believe there is a case for adopting the strict relation $\bar{w}_e < \bar{w}_i$ in the macroscopic dynamic model: a simple approximation to the percentage synapse distribution is to determine the total number of excitatory N_e and inhibitory N_i connections in a structural model and set $\bar{w}_i = -\bar{w}_e \times N_e/N_i$, thereby setting the total absolute weight for excitatory and inhibitory units to be equal. For the models described here, $\bar{w}_i = -0.2 \times 1176/573 = -0.41$.

Simulation of a macroscopic model with such an architecture shows that the cluster structure can implement soft selection (Figure 3b) - that is, simultaneous selection of more than one action. Some thresholding of output would be required to implement hard selection - a true WTA algorithm - a threshold possibly set by the amount of cluster output required to sufficiently activate their targets neurons in the cranial nerve nuclei and spine. The outputs for this simulation are, roughly, just the ratio of the corresponding inputs, which reduces the medial RF architecture to a simple relay system. Removing the inter-cluster connections to the projection neurons, by setting all $A_{jk} = 0$, leaves only the inter-cluster projections to inter-neurons and, thus, would seem more able to implement a WTA algorithm. However, simulation of this altered model shows that it does not implements a WTA algorithm either - the output of the clusters are little different from their input values (Figure 3c).

The existence of abundant long-range connections between projection neurons is not in doubt, and thus such an architecture cannot exist in the medial RF. Moreover, the presence or absence of the long-range connections appears to have little impact on the medial RF's ability to act as a selection mechanism if each cluster is assumed to represent a single action. Therefore, we are left to consider what purpose the long-range inter-cluster projection neuron connections have.

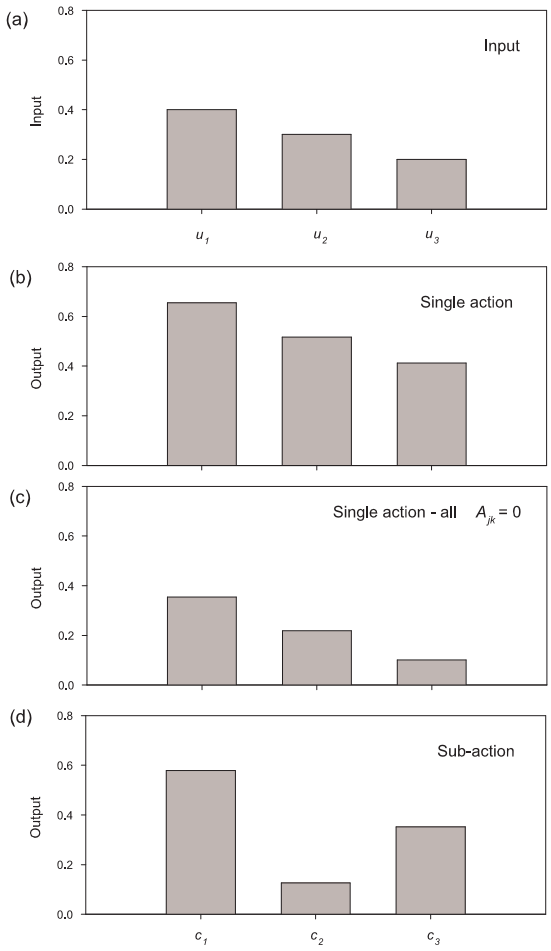


Figure 3: Example simulation results for configurations of a three cluster model. (a) Input values $\mathbf{u}$ to each cluster. (b-d) Cluster outputs: (b) a single-action configuration, with inter-cluster connections between projection neurons, does not act as a winner-takes-all (WTA) algorithm, but merely acts as an amplified relay of the inputs; (c) a single-action configuration, without inter-cluster connections between projection neurons, does not implement a WTA algorithm either; (d) a sub-action configuration, in which activation of cluster 1 (c_1) results in concurrent recruitment of cluster 3, and inhibition of cluster 2.

Sub-action configuration

It is useful to remember that the spatially-uniform model is just a statistical average - it is possible that some cluster-to-cluster projections preferentially target the inter-neuron populations, while others preferentially target the projection neurons. Thus, the output of a single cluster may simultaneously inhibit and excite different target clusters. The output of a

cluster in this scenario is taken to activate a sub-action, that is, a component part of a coherent behavior. Excitation of a target cluster corresponds to recruitment of a compatible, perhaps essential, sub-action; conversely, inhibition of a target cluster corresponds to the prevention of an incompatible, perhaps dangerous, sub-action. An example of this configuration in the same three cluster model is shown in Figure 2b. To generate this configuration, we take the previous $\mathbf{A}$ and $\mathbf{C}$ matrices - the mean numbers of inter-cluster connections to projection and inter-neurons, respectively - and set the appropriate connections to zero (see Appendix A). In simulation, the resulting cluster outputs (Figure 3d) show that the outputs of both clusters 1 and 3 have exceeded the value of their inputs, and both have considerably greater output than cluster 2 (which has a much reduced output compared to its input). Thus, in this configuration, the output pattern means that sub-actions 1 and 3 are activated, and sub-action 2 is not.

Having demonstrated that the sub-action configuration works in principle, we turn now to a preliminary assessment of its robustness over a range of inputs. The configuration depicted in Figure 2b supports just two actions, one signalled by the sufficient output of both clusters 1 and 3, and the other by the sufficient output of cluster 2. In this initial assessment, we deem sufficient output to mean that the outputs of the required clusters exceeds those of all the other clusters - that is, the selection of a sub-action is based solely on the ordering of the output values. Thus, given any set of inputs $\mathbf{u}$, we may define two correct output states:

1. if the outputs are ordered such that $(c_1 > c_2) \wedge (c_3 > c_2)$ then action 1 is correctly selected if and only if the input relationship is $(u_1 \geq u_2) \vee (u_3 \geq u_2)$,
2. if outputs are ordered such that $(c_2 > c_3) \wedge (c_2 > c_1)$ then action 2 is correctly selected if and only if the input relationship is $(u_2 \geq u_1) \wedge (u_2 \geq u_3)$,

where $\wedge$ means propositional conjunction and $\vee$ means propositional disjunction. All other alternatives are deemed to be incorrect selections (the example in Figure 3d fulfills output state 1, and is, therefore, a correct selection). We note that these are hard definitions of correct selection, in particular that both sub-actions that comprise action 1 must be selected together at all times (other interpretations, such as the correct selection of individual sub-actions, given appropriate inputs, will be considered in future work).

To assess the robustness of sub-action selection we simulated the model just described, varying each element of input vector $\mathbf{u}$ over the interval [0,1] in steps of 0.05, making a total of 1331 simulations. For each input vector, the output vector $\mathbf{c}$ was assessed to determine whether it signaled correct or incorrect selection, as defined above. We find that the majority of input vectors (75%) result in correct selection, as shown in the state space plots of Figure 4, and thus the sub-action selection is robust over a wide-range of inputs. The incorrect selections occurred for the input vectors for which either all the elements were roughly equal, or at least element u_2 and one other was, with the third element being close to zero. Thus, this simple model of a configuration of the medial RF's anatomy lacks a mechanism for resolving selection competitions between closely-matched inputs.

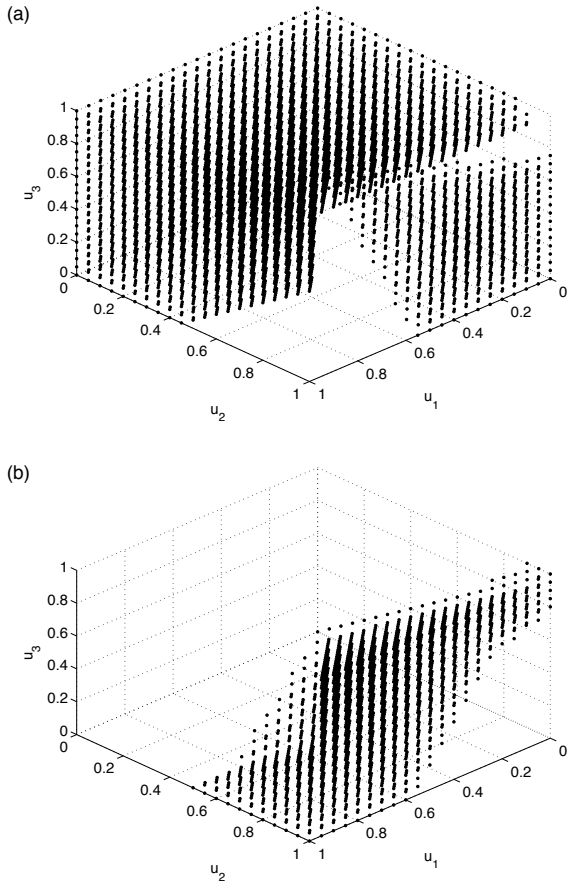


Figure 4: Output states of the sub-action configuration. (a) Correct selections. These occurred following the majority of inputs. (b) Incorrect selections. These occurred around the input values for which u_2 was roughly equal to either or both of u_1 and u_3 - where only one was roughly equal, the other of that pair was closer to zero.

5 Discussion

In this paper, we have extended our quantitative structural model of medial RF to incorporate sensory input, reviewed evidence which suggests an organisation of both inputs and outputs at a macroscopic organisational level we have previously dubbed the "cluster", and discussed how, given these biological constraints, the medial RF may be configured as the brainstem's action selection system. The example simulations demonstrate that a sub-action configuration, in which strongly activated clusters recruit clusters representing compatible sub-actions, provides a functional role for the abundant excitatory inter-cluster connections between projection neurons, as opposed to the single-action configuration, for which no role for those same connections was evident. In addition, the sub-action configuration was able to select appropriate actions over a wide range of input vectors. However, the simple model explored here was unable to appropriately resolve selection competitions between closely matched inputs. This may not be a fault of the model: it may be the case

that the isolated medial RF is actually unable to resolve such competition. Thus, it may be a contributing factor to the evolution of more complex action selection systems such as the basal ganglia.

To extend the work described here, we have two main threads. First, the addition of further features of the medial RF to the structural model and its dynamic instantiation may provide mechanisms which are able to resolve competitions between closely matched inputs. The existence of synapses for neuro-modulators, particularly serotonin, noradrenaline, and acetylcholine [Jones, 1995], means that neural activity within the medial RF could be up- or down-regulated according to local concentrations of these neurochemicals. Specifically, dependent on the neuro-modulator synapse type, input to this region could be locally enhanced or attenuated, providing a direct method for differentiating the responses to closely matched inputs. Second, to determine how the medial RF structure may support the relatively complex actions observed in the decerebrate animal, we will optimise the model's structure using a genetic algorithm in an embodied robotic task, following a methodology we have previously developed [Humphries *et al.*, a]. Our model of the basal ganglia [Gurney *et al.*, 2001] will also be optimised using the same task, so that we will be able to assess the relative merits of the two putative action selection substrates.

The sub-action configuration is given substantial support by data from a progressive decerebration study in which the grooming behavior of rats was assessed following a series of lesions descending from the midbrain to the junction between the pons and medulla [Berridge, 1989]. The intact brainstem is sufficient to support the entire sequence of actions which comprises the grooming syntax. Component actions of the grooming syntax, corresponding to what we've here called sub-actions, are disabled in an incremental fashion with descending decerebration, and thus we know that: (a) there is no single locus for the action of grooming in the brainstem and (b) there is not a widely-distributed representation of each grooming sub-action in the brainstem - for if there was, then descending decerebration should result in partial degradation of all components of the syntax, as their neural networks are damaged: instead, each component was either performed or entirely absent. Thus, there is good evidence for the existence of discrete, localised sub-action representations in the brainstem, corresponding to the structural and dynamical properties discussed for our medial RF model.

Acknowledgments

This work was supported by EPSRC grant GR/R95722/01.

A Model descriptions

The structural values for the particular instantiation of the structural model explored in simulation are:

$$\mathbf{A} = \begin{bmatrix} 0 & 1.61 & 1.57 \\ 2.15 & 0 & 2.09 \\ 2.03 & 2.49 & 0 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 1.7 \\ 2.14 \\ 2.03 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1.5 & 1.5 \\ 1.7 & 0 & 2.5 \\ 1.85 & 2.05 & 0 \end{bmatrix} \qquad \mathbf{d} = \begin{bmatrix} 1.8 \\ 1.55 \\ 2.1 \end{bmatrix}.$$

For the sub-action configuration, $\mathbf{A}$ and $\mathbf{C}$ are altered to match the connection pattern shown in Figure 2b, thus

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1.57 \\ 2.15 & 0 & 0 \\ 2.03 & 0 & 0 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} 0 & 1.5 & 0 \\ 1.7 & 0 & 2.5 \\ 0 & 2.05 & 0 \end{bmatrix}.$$

References

[Berntson and Micco, 1976] G. G. Berntson and D. J. Micco. Organization of brainstem behavioral systems. *Brain Research Bulletin*, 1:471–483, 1976.

[Berridge, 1989] K. C. Berridge. Progressive degradation of serial grooming chains by descending decerebration. *Behavioural and Brain Research*, 33(3):241–253, 1989.

[Drew and Rossignol, 1990] T. Drew and S. Rossignol. Functional organization within the medullary reticular formation of intact unanesthetized cat. I. Movements evoked by microstimulation. *Journal of Neurophysiology*, 64(3):767–781, 1990.

[Eccles *et al.*, 1976] J. C. Eccles, R. A. Nicoll, T. Rantucci, H. Taborikova, and T. J. Willey. Topographic studies on medial reticular nucleus. *Journal of Neurophysiology*, 39(1):109–118, 1976.

[Grantyn *et al.*, 1987] A. Grantyn, V. Ong-Meang Jacques, and A. Berthoz. Reticulo-spinal neurons participating in the control of synergic eye and head movements during orienting in the cat. II. Morphological properties as revealed by intra-axonal injections of horseradish peroxidase. *Experimental Brain Research*, 66(2):355–377, 1987.

[Grill and Kaplan, 2002] H. J. Grill and J. M. Kaplan. The neuroanatomical axis for control of energy balance. *Frontiers in Neuroendocrinology*, 23(1):2–40, 2002.

[Gurney *et al.*, 2001] K. Gurney, T. J. Prescott, and P. Redgrave. A computational model of action selection in the basal ganglia II: Analysis and simulation of behaviour. *Biological Cybernetics*, 85:411–423, 2001.

[Humphries *et al.*, a] M. D. Humphries, K. Gurney, and T. J. Prescott. Is there an integrative center in the vertebrate brainstem? a robotic evaluation of a model of the reticular formation viewed as an action selection device. *Adaptive Behavior*. in press.

[Humphries *et al.*, b] M. D. Humphries, K. Gurney, and T. J. Prescott. Small world and scale-free properties of the brainstem reticular formation. pre-print.

[Jones *et al.*, 1991] B. E. Jones, C. J. Holmes, E. Rodriguez-Veiga, and L. Mainville. GABA-synthesizing neurons in the medulla: their relationship to serotonin-containing and spinally projecting neurons in the rat. *Journal of Comparative Neurology*, 313(2):349–367, 1991.

[Jones, 1995] B. E. Jones. Reticular formation: Cytoarchitecture, transmitters, and projections. In G. Paxinos, editor, *The Rat Nervous System, 2nd Edition*, pages 155–171. New York: Academic Press, 1995.

[Kilmer *et al.*, 1969] W. L. Kilmer, W. S. McCulloch, and J. Blum. A model of the vertebrate central command system. *International Journal of Man-Machine Studies*, 1:279–309, 1969.

[Langhorst *et al.*, 1996] P. Langhorst, B. G. Schulz, H. Seller, and H. P. Koepchen. Convergence of visceral and somatic afferents on single neurones in the reticular formation of the lower brain stem in dogs. *Journal of the Autonomic Nervous System*, 57(3):149–157, 1996.

[Latham *et al.*, 2000] P. E. Latham, B. J. Richmond, P. G. Nelson, and S. Nirenberg. Intrinsic dynamics in neuronal networks. I. Theory. *Journal of Neurophysiology*, 83(2):808–827, 2000.

[Mink and Thach, 1993] J. W. Mink and W. T. Thach. Basal ganglia intrinsic circuits and their role in behavior. *Current Opinion in Neurobiology*, 3:950–957, 1993.

[Monteiro *et al.*, 2002] L. H. Monteiro, M. A. Bussab, and J. G. Chaui Berlink. Analytical results on a Wilson-Cowan neuronal network modified model. *Journal of Theoretical Biology*, 219(1):83–91, 2002.

[Prescott *et al.*, 1999] T. J. Prescott, P. Redgrave, and K. Gurney. Layered control architectures in robots and vertebrates. *Adaptive Behavior*, 7:99–127, 1999.

[Redgrave *et al.*, 1999] P. Redgrave, T. J. Prescott, and K. Gurney. The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience*, 89(4):1009–1023, 1999.

[Scheibel, 1984] A. B. Scheibel. The brainstem reticular core and sensory function. In J. M. Brookhart and V. B. Mountcastle, editors, *Handbook of Physiology. Section 1: The Nervous System*, pages 213–256. American Physiological Society: Bethesda, Maryland, 1984.

[Schulz *et al.*, 1983] B. Schulz, M. Lambertz, G. Schulz, and P. Langhorst. Reticular formation of the lower brainstem. A common system for cardiorespiratory and somatomotor functions: discharge patterns of neighboring neurons influenced by somatosensory afferents. *Journal of the Autonomic Nervous System*, 9(2-3):433–449, 1983.

[Schuz, 1995] A. Schuz. Neuroanatomy in a computational perspective. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 622–626. MIT Press, Cambridge, MA, 1995.

[Siegel and Tomaszewski, 1983] J. M. Siegel and K. S. Tomaszewski. Behavioral organization of reticular formation: studies in the unrestrained cat. I. Cells related to axial, limb, eye, and other movements. *Journal of Neurophysiology*, 50(3):696–716, 1983.

[Siegel *et al.*, 1981] J. M. Siegel, R. Nienhuis, R. L. Wheeler, D. J. McGinty, and R. M. Harper. Discharge pattern of reticular formation unit pairs in waking and REM sleep. *Experimental Neurology*, 74(3):875–891, 1981.

[Siegel, 1979] J. M. Siegel. Behavioral functions of the reticular formation. *Brain Research Reviews*, 1:69–105, 1979.

Contracting model of the basal ganglia*

Benoît Girard[1], **Nicolas Tabareau**[1], **Jean-Jacques Slotine**[2] **and Alain Berthoz**[1]
1. Laboratoire de Physiologie de la Perception et de l'Action, CNRS - Collège de France
11 place Marcelin Berthelot, 75231 Paris Cedex 05, France.
2. Nonlinear Systems Laboratory, Massachusetts Institute of Technology
Cambridge, Massachusetts, 02139, USA

Abstract

It is thought that one role of the basal ganglia is to constitute the neural substrate of action selection. We propose here a modification of the action selection model of the basal ganglia of (Gurney et al., 2001a,b) so as to improve its dynamical features. The dynamic behaviour of this new model is assessed by using the theoretical tool of contraction analysis. We simulate the model in the standard test defined in (Gurney et al., 2001b) and also show that it performs perfect selection when presented a thousand successive random entries. From a biomimetical point of view, our model takes into account a usually neglected projection from GPe to the striatum, which enhances its efficiency.

Keywords: contraction analysis, action selection, basal ganglia, computational model

1 Introduction

The basal ganglia are a set of interconnected subcortical nuclei, involved in numerous processes, from motor functions to cognitive ones (Mink, 1996; Middleton and Strick, 1994). Their role is interpreted as a generic selection circuit, and they thus have been proposed to constitute the neural substrate of action selection (Mink, 1996; Krotopov and Etlinger, 1999; Redgrave et al., 1999).

The basal ganglia are included in cortico-basal ganglia-thalamo-cortical loops, five main loops have been identified in primates (Alexander et al., 1986, 1990; Kimura and Graybiel, 1995): motor, oculomotor, prefrontal (two of them) and limbic loops. Within each of these loops, the basal ganglia circuitry is organised in interacting channels, among which selection occurs. The output nuclei of the basal ganglia are tonically active and inhibitory, and thus maintain their targets under sustained inhibition. Selection occurs *via* disinhibition (Chevalier and Deniau, 1990): the removal of the inhibition exerted by one channel on its specific target circuit allows the activation of that circuit. Concerning action selection, the basal ganglia channels are thought to be associated to basic competing actions. Given sensory and motivational inputs, the basal ganglia are thus supposed to arbitrate among these actions and to allow the activation of the winner by disinhibiting the corresponding motor circuits.

Numerous computational models of the BG have been proposed in the past (Gillies and Arbruthnott, 2000, for a review) in order to explain the operation of this disinhibition process, the most recent and complete model –in terms of anatomically identified connections accounted– is the GPR model proposed by Gurney et al. (2001a,b). Beyond its generic selection properties, explored in (Gurney et al., 2001b), the efficiency of the GPR as an action selection device has been tested in both robotic and simulated animats solving various tasks, involving execution of behavioural sequences, survival and navigation (Montes-Gonzalez et al., 2000; Girard et al., 2003, 2005).

The properties of the GPR were analytically studied at equilibrium, however the stability of this equilibrium (and thus the possibility to reach it) was not assessed. We propose to use contraction analysis (Lohmiller and Slotine, 1998) –a theoretical tool to study the dynamic behaviour of non-linear systems– in order to build a new model of the basal ganglia whose stability can be formally established. By using recent data (Parent et al., 2000) concerning the projections of a basal ganglia nucleus (the external part of the globus pallidus), we improve the quality of its selection with regards to GPR and then test this improvement in simulation. Finally, we discuss the remaining biomimetic limitations of the proposed model.

2 Nonlinear Contraction Analysis

Basically, a nonlinear time-varying dynamic system will be called *contracting* if initial conditions or temporary disturbances are forgotten exponentially fast, i.e., if trajectories of the perturbed system return to their nominal behaviour with an exponential convergence rate. This is an extension of the well-known *stability* analysis for linear systems with the great advantage that relatively simple conditions can still be given for this stability-like property to be verified, and furthermore that this property is preserved through basic system combinations. We also want to stress that assuming that a system is contracting, we only have to find a particular stable trajectory to be sure that the system will eventually tend to this trajectory. It is thus a way to analyse the dynamic behaviour of a model without linearised approximation.

*The support of the BIBA project funded by the European Community, grant IST-2001-32115 is acknowledged.

2.1 The basic brick

In this section, we summarise the variational formulation of contraction analysis of (Lohmiller and Slotine, 1998), to which the reader is referred for more details. It is a way to prove the contraction of a whole system by analysing the properties of its Jacobian only. This can be seen as the basic brick of the theory, as in next sections we will often study the contraction of small components of the system and then deduce the global contraction of the system using combination rules (see section 2.2).

Consider a n-dimensional time-varying system of the form:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), t) \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}_+$ and $\mathbf{f}$ is $n \times 1$ non-linear vector function which is assumed to be real and smooth in the sense that all required derivatives exist and are continuous. This equation may also represent the closed-loop dynamic of a neural network model of a brain structure.

We now restate the main result of contraction analysis, see (Lohmiller and Slotine, 1998) for details and proof.

Theorem 1 *Consider the continuous-time system (1). If there exists a uniformly positive definite metric*

$$\mathbf{M}(\mathbf{x}, t) = \mathbf{\Theta}(\mathbf{x}, t)^T \mathbf{\Theta}(\mathbf{x}, t)$$

such that the generalised Jacobian

$$\mathbf{F} = (\dot{\mathbf{\Theta}} + \mathbf{\Theta}\mathbf{J})\mathbf{\Theta}^{-1}$$

is uniformly negative definite, then all the all system trajectories converge exponentially to a single trajectory with convergence rate $|\lambda_{max}|$, where λ_{max} is the largest eigenvalue of the symmetric part of $\mathbf{F}$. The system is said to be contracting.

Remark. In many cases, if the system is not properly defined, the expected metric may be hard to find. Most often, it is possible to fall into a standard combination of contracting systems just by rearranging the order of variables considered whereas the original definition of the system did not stress contraction properties.

2.2 Combination of contracting systems

We now present standard results on combination of contracting systems which will help us in showing that our model is contracting by analysing first contraction of each nucleus on one side and then their relative combination.

Hierarchies

The most useful combination is the hierarchical one. Consider a virtual dynamic of the form

$$\frac{d}{dt}\begin{pmatrix} \delta\mathbf{z}_1 \\ \delta\mathbf{z}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{F}_{11} & \mathbf{0} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{pmatrix}\begin{pmatrix} \delta\mathbf{z}_1 \\ \delta\mathbf{z}_2 \end{pmatrix}$$

The first equation does not depend on the second, so that exponential convergence of the whole system can be guaranteed (Lohmiller and Slotine, 1998). The results can be applied recursively to combinations of arbitrary size.

Feedback Combination

Consider two contracting systems and an arbitrary feedback connection between them (Slotine, 2003). The overall virtual dynamics can be written

$$\frac{d}{dt}\begin{pmatrix} \delta\mathbf{z}_1 \\ \delta\mathbf{z}_2 \end{pmatrix} = \mathbf{F}\begin{pmatrix} \delta\mathbf{z}_1 \\ \delta\mathbf{z}_2 \end{pmatrix}$$

Compute the symmetric part of $\mathbf{F}$, in the form

$$\frac{1}{2}\left(\mathbf{F} + \mathbf{F}^T\right) = \begin{pmatrix} \mathbf{F}_{1s} & \mathbf{G}_s \\ \mathbf{G}_s^T & \mathbf{F}_{2s} \end{pmatrix}$$

where by hypothesis the matrices $\mathbf{F}_{is}$ are uniformly negative definite. Then $\mathbf{F}$ is uniformly negative definite if and only if $\mathbf{F}_{2s} < \mathbf{G}_s^T \mathbf{F}_{1s}^{-1} \mathbf{G}_s$, a standard result from matrix algebra (Horn and Johnson, 1985). Thus, a sufficient condition for contraction of the overall system is that

$$\sigma^2(\mathbf{G}_s) < \lambda(\mathbf{F}_1)\,\lambda(\mathbf{F}_2) \qquad \text{uniformly } \forall \mathbf{x},\ \forall \mathrm{t} \geq 0$$

where $\lambda(\mathbf{F}_i)$ is the contraction rate of $\mathbf{F}_i$ and $\sigma(\mathbf{G}_s)$ is the largest singular value of $\mathbf{G}_s$. Again, the results can be applied recursively to combinations of arbitrary size.

Contraction analysis on convex regions

Consider a contracting system $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, t)$ maintained in a convex region $\mathbf{\Omega}$ (i.e. a region $\mathbf{\Omega}$ in which any shortest connecting line (geodesic) $\int_{\mathbf{x}_1}^{\mathbf{x}_2} \|\delta\mathbf{x}\|$ between two arbitrary points $\mathbf{x}_1$ and $\mathbf{x}_2$ in $\mathbf{\Omega}$ is completely contained in $\mathbf{\Omega}$). Then all trajectories in $\mathbf{\Omega}$ converge exponentially to a single trajectory (Lohmiller and Slotine, 2000). Furthermore, the contraction rate can only be sped up by the convex constraint.

2.3 Our basic contracting system : the leaky integrator

In our model of basal ganglia, we will use leaky integrator models of neurons. The following equations describe the behaviour of our neurons where τ is a time constant $a(t)$ is the activation, $y(t)$ is the output, $I(t)$ represents the input of the neuron, and f is a continuous function which maintains the output in an interval.

$$\begin{cases} \tau\dot{a}(t) &= -a(t) + I(t) \\ y &= f(a) \end{cases}$$

This kind of neuron is basically contracting since its Jacobian is $-\frac{1}{\tau}$ and the interval defined by the transfer function is a particular convex region.

In the rest of this paper, we will use the family of functions $f_{\varepsilon,max}$:

$$\begin{cases} 0 & \text{if} \quad x \leq \varepsilon \\ x - \varepsilon & \text{if} \quad \varepsilon \leq x \leq max + \varepsilon \\ max & \text{else} \end{cases} \tag{2}$$

3 Model description

The basic architecture of our model is very similar to the GPR (fig 1). We use the same leaky-integrator model of neurons as building blocks, each BG channel in each nucleus being represented by one such neuron. The input of the system is a

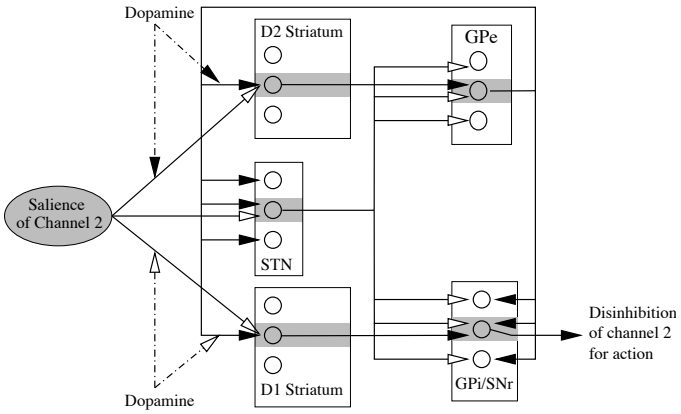


Figure 1: Basal ganglia model. Nuclei are represented by boxes, each circle in these nuclei represents an artificial leaky-integrator neuron. On this diagram, three channels are competing for selection, represented by the three neurons in each nucleus. The second channel is represented by grey shading. For clarity, the projections from the second channel neurons only are represented, they are identical for the other channels. White arrowheads represent excitations and black arrowheads, inhibitions. D1 and D2: neurons of the striatum with two respective types of dopamine receptors; STN: subthalamic nucleus; GPe: external segment of the globus pallidus; GPi/SNr: internal segment of the globus pallidus and substantia nigra pars reticulata.

vector of saliences, representing the propensity of each behaviour to be selected. Each behaviour in competition is associated to a specific channel and can be executed if and only if its level of inhibition decreases below a fixed threshold θ.

An important difference between the GPR and our model is the nuclei targeted by the external part of the globus pallidus (GPe) and the nature of these projections. The GPe projects to the subthalamic nucleus (STN), the internal part of the globus pallidus (GPi) and the substantia nigra pars reticulata (SNr), but also to the striatum. Our model includes the striatum projections, which have been documented (Staines et al., 1981; Kita et al., 1999) but excluded from previous models. Moreover, the striatal terminals target the dendritic trees, while pallidal, nigral and subthalamic terminals form perineuronal nets around the soma of the targeted neurons (Sato et al., 2000). This specific organisation allows GPe neurons to influence large sets of neurons in GPi, SNr and STN (Parent et al., 2000), thus the sum of the activity of all GPe channels influences the activity of STN and GPi/SNr neurons (eqn. 5 and 7), while there is a simple channel-to-channel projection to the striatum (eqn. 3 and 4).

The striatum is one of the two input nuclei of the BG, mainly composed of GABAergic (inhibitory) medium spiny neurons. As in the GPR model, we distinguish the neurons with D1 and D2 dopamine receptors and modulate the input generated in the dendritic tree by λ, which here encompasses salience and GPe projections. Lateral inhibitions are also implemented, but their weights w_{LatD1} and w_{LatD2} is kept within the limits set the contraction analysis (see section 4.1). The input to each neuron i of the D1 and D2 sub parts of the striatum is therefore defined as follows (N being the number of channels):

$$I_i^{D1} = (1+\lambda)(S_i - w_{GPe}^{D1} y_i^{GPe}) - w_{LatD1} \sum_{\substack{j=1 \\ j\neq i}}^{N} I_j^{D1} \quad (3)$$

$$I_i^{D2} = (1-\lambda)(S_i - w_{GPe}^{D2} y_i^{GPe}) - w_{LatD2} \sum_{\substack{j=1 \\ j\neq i}}^{N} I_j^{D2} \quad (4)$$

The up-state/down-state of the striatal medium spiny neurons is modelled, as in (Gurney et al., 2001b), by activation thresholds ε_{D1} and ε_{D2} under which the neurons remain silent.

The sub-thalamic nucleus (STN) is the second input of the basal ganglia and receives also projections from the GPe. Its glutamatergic neurons have an excitatory effect and project to the GPe and GPi. The resulting input of the STN neuron is given by:

$$I_i^{STN} = S_i - w_{GPe}^{STN} \sum_{j=1}^{N} y_j^{GPe} \quad (5)$$

The tonic activity of the nucleus is modelled by a negative threshold of the transfer function ε_{STN}.

The GPe is inhibitory nucleus, similarly as in the GPR, it receives channel-to-channel afferents from the striatum and a diffuse excitation from the STN:

$$I_i^{GPe} = -w_{D2}^{GPe} y_i^{D2} + w_{STN}^{GPe} \sum_{j=1}^{N} y_j^{STN} \quad (6)$$

The GPi and SNr are the inhibitory output nuclei of the BG, which keep their targets under inhibition unless a channel is selected. They receive channel-to-channel projections from the D1 striatum and diffuse projections from the STN and the GPe:

$$\begin{aligned} I_i^{GPi} = & - w_{D1}^{GPi} y_i^{D1} + w_{STN}^{GPi} \sum_{j=1}^{N} y_i^{STN} \\ & - w_{GPe}^{GPi} \sum_{j=1}^{N} y_j^{GPe} \end{aligned} \quad (7)$$

This model keeps the basic off-centre on-surround selecting structure, duplicated in the D1-STN-GPi/SNr and D2-STN-GPe sub-circuits, of the GPR. However, the channel specific feedback from the GPe to the Striatum helps sharpening the selection by favouring the channel with the highest salience in D1 and D2. Moreover, the global GPe inhibition on the GPi/SNr synergetically interacts with the STN excitation in order to limit the amplitude of variation of the inhibition of the unselected channels.

4 Mathematical results

We first analyse the contraction of the GPR model before showing under which weighting constraints our model is contracting and which sufficient salience input conditions allow it to perform "perfect selection" (output inhibition of selected channels equal to 0).

4.1 Contraction analysis of the GPR model

While it is difficult to refute contraction of a system as the metric in which it is contracting is not given *a priori*, we can study contraction in particular metrics for the sake of finding a contra-example which will demonstrate the non-contracting behaviour of the system.

First, remark that lateral connections on striatum (D_1 and D_2) make the model non-contracting in the identity metric when the weight of inhibition $w_{Lat} \geq 1$. Indeed, by computing directly the eigenvalues of the Jacobian

$$J = \begin{pmatrix} -1 & -w_{Lat} & . & . & -w_{Lat} \\ -w_{Lat} & -1 & . & . & . \\ . & . & . & . & . \\ . & . & . & . & -w_{Lat} \\ -w_{Lat} & . & . & -w_{Lat} & -1 \end{pmatrix}$$

we have $\lambda_{max} \leq -1 + w_{Lat}$. Unsurprisingly, when $w_{Lat} = 1$ the system has multiple points of stability and thus the model is not contracting in any metric.

A typical example of multiple points of stability occurs when two channels, say i and j, have the same highest salience S_{max} for input. We then have a continuum of possible stable points in D1 and D2 covering the segment $a_i + a_j = S_{max}$ with $a_i, a_j \geq 0$, while all the other channels being fully inhibited.

Such a situation occurs when reproducing the basic selection test proposed in (Gurney et al., 2001b). In this five-steps test (fig. 2), no channels are excited during the first one, and none of them is thus selected; then during the second one, the salience of channel 1 is increased and this channel is consequently selected; during the third one, channel 2 is provided a larger salience than channel 1, channel 1 is thus inhibited and channel 2 selected; in the fourth one, the salience of channel 1 is increased to a value equal to the salience of channel 2, *channel 1 is however not selected while channel 2 remains selected*; finally the salience of channel 1 is decreased to its initial level. Such a drawback can only be solved by reducing w_{Lat} to a value strictly inferior to 1.

Second, suppose w_{Lat} is set under 1 to avoid this specific problem, it remains to show that the GPe/STN loop is contracting. Using the feedback analysis with a scaling metric that dilates the states space of the second system involved (a key tool in the study of many feedbacks)

$$\mathbf{M} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \alpha\mathbf{I} \end{pmatrix},\ \alpha > 0$$

makes us compute the maximum singular value of G_s (see section 2.2):

$$\sigma(G_s) = max(\frac{\alpha}{2}, \frac{1}{2}(-\alpha w_{GPe}^{STN} + \frac{N}{\alpha} w_{STN}^{GPe}))$$

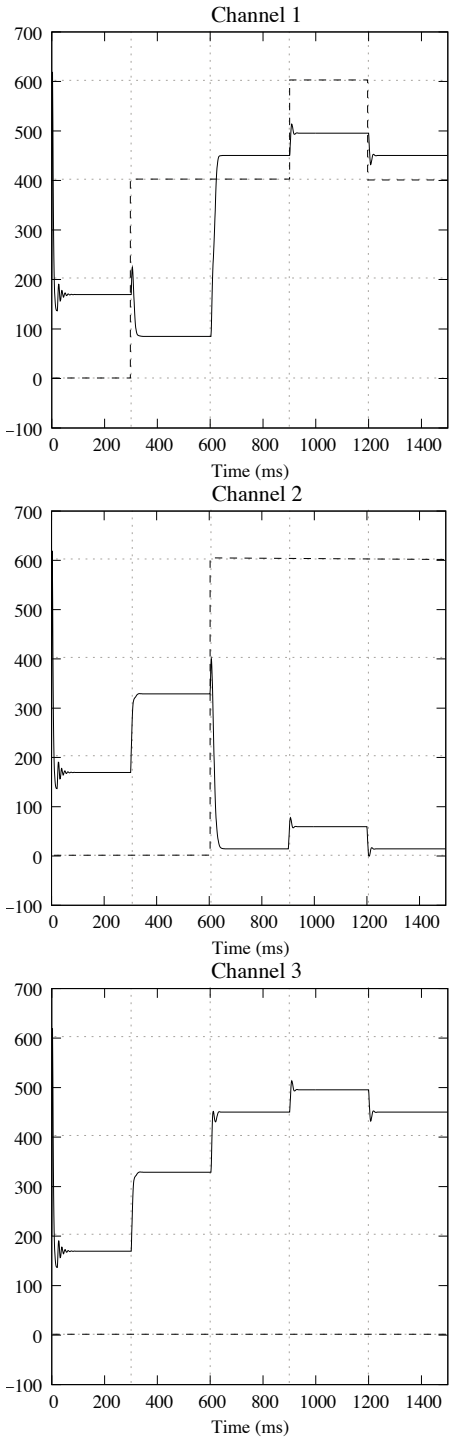


Figure 2: Simulation results (GPi/SNr inhibitory output) for the first three channels of a 6-channels system, using the Gurney et al. (2001b) test on the GPR model. During the period $900ms < t < 1200ms$, channels 1 and 2 have the same input saliences, and channel 2 only is selected. Dashed lines represent the input salience of the channel and solid lines represent the output of the channel.

which gives rise to the following condition on N :

$$N < \frac{4}{w_{STN}^{GPe}}(1 + w_{GPe}^{STN})$$

Analysed in the scaling metric, the contraction of the GPR is proven when N remains below this bound, which corresponds to $N < 6$ with the parameters used in (Gurney et al., 2001b). This does not strictly demonstrate that the GPR model with lateral striatal inhibitons lower than 1 is not contracting for $N \geq 6$, as there might be another metric in which the analysis would give a contraction result with a different dependence on, or even an independence from, N. It however suggests that, even if the result is not conclusive, the conditions of contraction of the GPR model probably depend on N, this is the main motivation for proposing a model whose contraction is proven for less restrictive conditions.

4.2 Contraction of the model

The contraction of our model is demonstrated using the combination properties of contracting systems.

First, we see that every nucleus is trivially contracting with a rate $\frac{1}{\tau}$ as no lateral connection is allowed except for the D_i's which are contracting when $w_{LatDi} < 1$ with rate $\frac{1}{\tau}|-1+w_{LatDi}|$ (see section 4.1). Dealing with thresholds of the leaky-integrator transfer functions is transparent as it is just a particular case of contraction analysis on convex regions (see section 2.2).

Next, defining the system carefully leads to a hierarchical system of trivially contracting systems except for the loops between STN/GP_e and D_2/GP_e. Thus, we only have to master those loops thanks to the feedback combination analysis to guarantee contraction of the whole system.

STN/GPe

Thanks to our reformulation of the GPe to STN projections (diffuse rather than channel-to-channel), this loop is now contracting as it is a positive/negative feedback. In other word, considering the metric

$$M_1 = \begin{pmatrix} w_{GPe}^{STN}\,\mathbf{I} & \mathbf{0} \\ \mathbf{0} & w_{STN}^{GPe}\,\mathbf{I} \end{pmatrix}$$

leads to the generalised Jacobian

$$F = \begin{pmatrix} -\mathbf{I} & (w_{GPe}^{STN} w_{STN}^{GPe})^{\frac{1}{2}}\,\mathbf{1} \\ -(w_{GPe}^{STN} w_{STN}^{GPe})^{\frac{1}{2}}\,\mathbf{1} & -\mathbf{I} \end{pmatrix}$$

and the feedback thus disappears as the symmetrical of $\mathbf{F}$ is simply $-\mathbf{I}$.

D2/GPe

The feedback is of the form negative/negative feedback and thus we can just try to minimise the impact of the loop by taking the average of each negative feedback. This is realised by considering the metric

$$M_2 = \begin{pmatrix} w_{D2}^{GPe}\,\mathbf{I} & \mathbf{0} \\ \mathbf{0} & w_{GPe}^{D2}\,\mathbf{I} \end{pmatrix}$$

which tells us that the system is contracting as long as

$$w_{D2}^{GPe} w_{GPe}^{D2} < -1 + w_{LatD2}$$

The last equation is obtained by using feedback analysis, see section 2.2 for more details.

4.3 Analytical results

As our model is contracting, we only have to find a particular solution to be sure that the system will eventually reach this solution. But, because this contracting system is autonomous (time-invariant), we know that this solution is an equilibrium (Slotine, 2003). Thus, it just remains to show that this equilibrium performs the awaited selection.

Naturally, as for GPR, we can show that our model is *order preserving* and that $\sum_{j=0}^{n} y_j^{STN}$ is bounded. But more interestingly, we can analytically study our model in the *ideal case* when the stable state is one active neuron only, say i_0, in D_2 and one inactive in GP_e (necessarily the same i_0). We call this situation *ideal case* as the selection is completely performed in the $D_2 - STN - GP_e$ loop and the rest of the model simply copies this selection.

Assuming that the salience input of the system leads to the this particular behaviour, we can obtain the following equations by solving the system of linear equations defined in section 3, using that $a = I$ for all neurons at equilibrium.

$$\sum_{j=1}^{N} y_j^{STN} = \frac{\sum_{y_j^{STN} \neq 0} (S_j + \varepsilon_{STN})}{1 + act(N-1) w_{STN}^{GPe} w_{GPe}^{STN}}$$

$$S_{i_0} \geq \varepsilon_{D_2} + \frac{w_{STN}^{GPe}}{w_{D2}^{GPe}} \sum_{j=1}^{N} y_j^{STN}$$

$$\begin{aligned} S_i \leq & \varepsilon_{D_2} + w_{LatD2}(S_{i_0} - \varepsilon_{D_2}) \\ & + w_{GPe}^{D2} w_{STN}^{GPe} \sum_{j=1}^{N} y_j^{STN} \qquad i \neq i_0 \end{aligned}$$

where act is the number of neurons of the STN whose activation is larger than ε_{STN}. Remark that when $(N-1) w_{STN}^{GPe} w_{GPe}^{STN} = 1$, $\sum_{j=0}^{n} y_j^{STN}$ computes essentially the mean of the active saliences.

Those equations give a range of saliences input for which the model reacts ideally, as its equilibrium corresponds to a "perfect selection", where the selected channel is completely disinhibited. Outside this range, the behaviour is more awkward as the whole system is involved in improving the partial selection made by the $D_2 - STN - GP_e$ loop. It might continue to perform "perfect selection", perform a less precise selection or behave differently, hence the simulation of section 5.2 in a wide set of input conditions.

5 Simulation results

Similarly to the simulations made by Gurney et al. (2001b), we used a 6-channel model. The parameters were set to the values summarised in table 1. w_{LatD1}, w_{LatD2}, w_{D2}^{GPe} and w_{GPe}^{D2} were set to values compatible with the constraints needed to ensure the contraction of the system (see 4.2). w_{GPe}^{D1} and w_{STN}^{GPi} were set to values identical to w_{GPe}^{D2} and w_{STN}^{GPe} respectively, for the sake of symmetry, whereas it is not mandatory with regards to contraction. Finally we set w_{D1}^{GPi} to 1 rather than to 0.7 (as w_{D2}^{GPe}) in order to favour strong selective inhibitions over GPi and thus "perfect selections".

The simulation was programmed in C++, using the simple Euler approximation for integration, with a time step of $1ms$.

Table 1: Parameters of the simulations.

w_{LatD1}	0.4	w_{D2}^{GPe}	0.7	τ	0.003s
w_{LatD2}	0.4	w_{D1}^{GPi}	1	λ	0.2
w_{GPe}^{D1}	1	w_{STN}^{GPe}	0.35	ε_{D1}	200
w_{GPe}^{D2}	1	w_{STN}^{GPi}	0.35	ε_{D2}	200
w_{GPe}^{STN}	0.35			ε_{STN}	-150
w_{GPe}^{GPi}	0.08				

5.1 Reproduction of GPR basic selection properties

We reproduced the selection experiment of Gurney et al. (2001b), were the system is submitted a sequence of five different salience vectors. As we bounded the activity of our neurons between 0 and 1000, while Gurney et al. had an upper limit of 1, we multiplied by 1000 the input saliences for this test. Each vector is submitted to the system during $0.3s$ before switching to the next one in the sequence (fig. 3).

First, all saliences are null, and the system stabilises in a situation where all channels are equally inhibited. Then, the first channel receives a 400 input salience which results in perfect disinhibition of this channel ($y_1^{GPi} = 0$) and increased inhibition of the others. When the second channel salience is set to 600, it becomes perfectly selected ($y_2^{GPi} = 0$) while the first one is rapidly inhibited to a level identical to the one of the four last channels. During the fourth step, the salience of the first channel is increased to 600, channels 1 and 2 are therefore simultaneously selected. Finally, during the last step of the test, the salience of channel 1 is reduced to 400, which is then rapidly inhibited while the selection of channel 2 is unaffected.

Our model passes this test in satisfactory manner, its results differ with the GPR in two ways. Firstly, it tends to select channels in a sharper manner than the GPR, as it always reaches "perfect selection" ($y_i^{GPi} = 0$). Secondly, the global level of inhibition in the unselected channels is subject to smaller variations, because of the regulatory effect of balance between the GPe global inhibition and the STN global excitation over the GPi.

5.2 1000 random vectors test

In order to test the ability of the model to perform "perfect selection" in a wide range of salience inputs and without any influence of its initial state (a property implied by contraction of the model), we fed a 6-channels system with a sequence of 1000 randomly drawn salience vectors successively. The saliences of each vectors are drawn uniformly in a 0 to 990 interval (discretisation step of 10), equal saliences are authorised within the same vector. Each vector is presented during $0.3s$, at the end of this period, the "perfect selection" of the channels with maximum salience is checked along with the presence of perfectly selected channels corresponding to other salience values. Then the next random vector is presented without resetting the system. This test was conducted with our model and with a GPR model for which w_{LatDi} was set to 0.8.

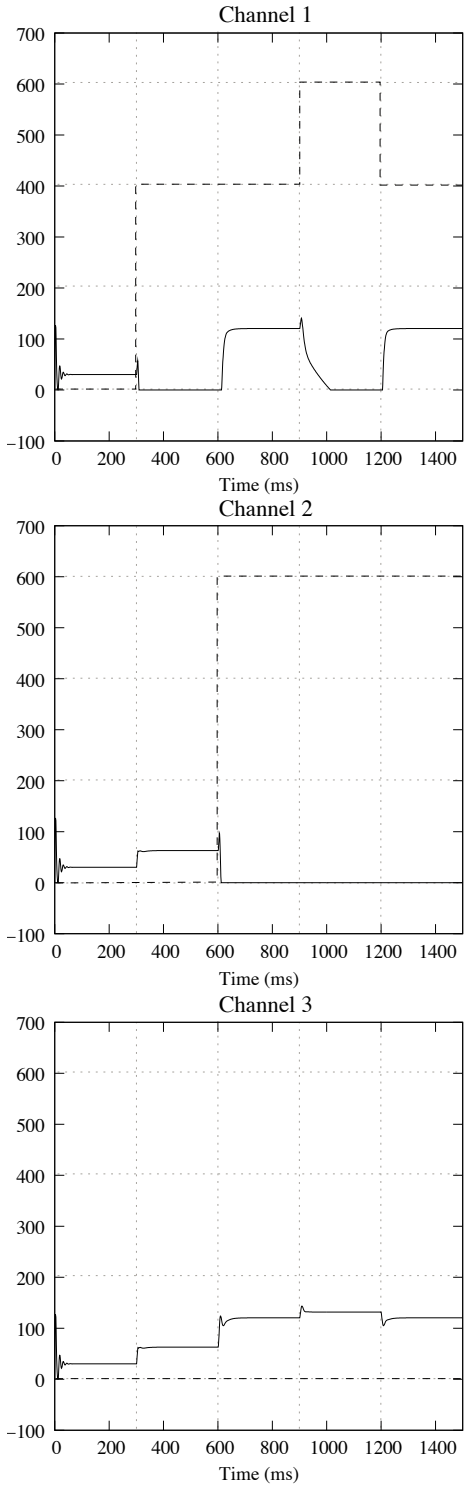


Figure 3: Simulation results (GPi/SNr inhibitory output) for the first three channels of a 6-channels system, using the Gurney et al. (2001b) test. Dashed lines represent the input salience of the channel and solid lines represent the output of the channel.

The first result of the test is that for our model, the "perfect selection" of the channels with maximum salience was not completed in only two cases out of thousand. This occurs when the maximum salience is too low to enable the activity in the striatal neurons to rise above the striatum thresholds $\varepsilon_{D1} = \varepsilon_{D2} = 200$ and is thus unable to elicit selection, an expected result as these thresholds are thought to filter low level saliences. Concerning the GPR, processing the same 1000-vectors sequence, "perfect selection" was not obtained in 54.6% of the cases, which is quite natural as the GPR is not designed to perform "perfect selection". The inhibitory output of the GPi/SNr of the GPR model is close 160 when the input salience vector is null, in which case no channel should be selected. We thus chose a value of the θ threshold equal to that maximum. In that case, there is no selection in 29.3% of the cases. It seems that in this range of salience input, our model selects winning channels more efficiently than the GPR.

The second result of the test is that the model has a nice property of contrast enhancement, as the maximum can be sorted out from its competitors even if they are quite close, generating a perfect selection of the former and a strong inhibition of the latter. Indeed, simultaneous selection of the channel with maximum salience with one of its competitors happens only in 7.2% of the cases. Moreover, this only happens when the maximal salience value is high ($\mu = 907.5$, $\sigma = 71.3$) and when the difference between the maximal salience and the salience of the supplementary selected channel is low (45 selections with a difference of 10, 24 with a difference of 20, 2 with 30 and 1 with 40). We may thus infer that the limit of discrimination between two saliences of our model is probably inferior to a few percents.

6 Discussion

We proposed a new computational model of the basal ganglia exploring how their intrinsic computations operate the physiologically observed "selection by disinhibition" (Chevalier and Deniau, 1990), which is thought to be a fundamental neural substrate of action selection in vertebrates (Redgrave et al., 1999). This model shares a lot of similarities with the previously proposed GPR model (Gurney et al., 2001b), as its selection ability relies on two off-centre on-surround subcircuits. However, it includes neglected connections from the GPe to the Striatum. Moreover, it distinguishes global projections of the GPe to the STN, GPi and SNr on the one hand and channel-to-channel ones to the Striatum on the other.

We theoretically studied the dynamic behaviour of the network and proved its stability by showing that it is contracting and has an equilibrium point, and thus always converges exponentially fast to this equilibrium. The independence of this contraction with regards to the number of channels results from the diffuse inhibitions from GPe to STN. We also showed that in an *ideal case*, implying conditions on the saliences values, this equilibrium corresponds to a *perfect selection* (where the channel corresponding to the highest salience is completely disinhibited and all others inhibited).

In order to test the selection efficiency of the model in a wider range of input conditions, we reproduced the basic selection test proposed by Gurney et al. (2001b) and, above all, evaluated the quality of selection when it is given a sequence of 1000 random salience vectors. In both cases, *perfect selection* was obtained, except in the rare cases where all the components of the salience vector are too low to elicit selection. Moreover, the selectivity of the model in the second test was better than the GPR.

We modelled the projections from GPe to striatum as having a channel-to-channel selectivity. However, in their study of five pallido-striatal neurons in rats, Bevan et al. (1998) showed that their primary target seems to be the GABAergic interneurons. First, given the limited extend of this study, we cannot exclude the possibility that GPe-striatum projections also concern striatum projection neurons. Second, the GABAergic interneurons inhibit the striatum projection neurons in a relatively diffuse manner, a regulatory effect that is different from but not opposed to our selective and direct projections: it controls the activity of the whole striatum and can thus affect the contrast of the selection. An alternate version of our model derived from these results should be tested.

We omitted two extra types of documented connections. First, the STN projects to the GPe, GPi and SNr but also to the striatum (Parent et al., 2000). Intriguingly, the population of STN neurons projecting to the striatum does not project to the other targets, while the other neurons project to at least two of the other target nuclei. We could not decipher the role of this striatum-projecting population and did not include it in the current model. Its unique targeting specificity suggests it could be functionally distinct from the other STN neurons. This possibility should be explored in future work. The other missing connections concerns the fact that D1 striatal neurons probably simultaneously project to the GPi/SNr and the GPe (Wu et al., 2000), and the fact that lateral inhibition exist in GPe and SNr (Park et al., 1982; Juraska et al., 1977; Deniau et al., 1982). These additional projections were added to the GPR in an improved implementation (Gurney et al., 2004), where the lateral inhibitions of the striatum were also removed. We should add these connections and proceed to a similar test with our model, knowing that the D1-GPe projections would create a new D1-GPe loop and generate an additional constraint on the weights to ensure contraction.

The GPe to striatum connections have the previously evoked functional advantage of enhancing the quality of the selection, by silencing the unselected striatal neurons. Interestingly, the striatum is known for being a relatively silent nucleus (Wilson, 1993), a property supposed to be induced by the specific up/down state behaviour of the striatal neurons. When using simple neuron models, like leaky-integrators, it is usually difficult to reproduce this with a threshold in the transfer function only: when many channels have a strong saliences input, all the corresponding striatal neurons tend to be activated. Our model suggests that in such a case, the GPe-striatum projections may contribute to silencing the striatum.

Finally, the basal ganglia are part of cortico-basal ganglia-thalamo-cortical loops and the quality of selection of the GPR model was improved by the addition of the thalamo-cortical components (Humphries and Gurney, 2002). We plan to extend our model in a similar manner while trying to preserve its contraction properties.

References

Alexander, G. E., Crutcher, M. D., and DeLong, M. R. (1990). Basal ganglia-thalamocortical circuits: Parallel substrates for motor, oculomotor, "prefrontal" and "limbic" functions. *Progress in Brain Research*, 85:119–146.

Alexander, G. E., DeLong, M. R., and Strick., P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, 9:357–381.

Bevan, M., Booth, P., Eaton, S., and Bolam, J. (1998). Selective innervation of neostriatal interneurons by a subclass of neurons in the globus pallidus of rats. *Journal of Neuroscience*, 18(22):9438–9452.

Chevalier, G. and Deniau, M. (1990). Disinhibition as a basic process of striatal functions. *Trends in Neurosciences*, 13:277–280.

Deniau, J.-M., Kitai, S., Donoghue, J., and Grofova, I. (1982). Neuronal interactions in the substancia nigra pars reticulata through axon collateral of the projection neurons. *Experimental Brain Research*, 47:105–113.

Gillies, A. and Arbruthnott, G. (2000). Computational models of the basal ganglia. *Movement Disorders*, 15(5):762–770.

Girard, B., Cuzin, V., Guillot, A., Gurney, K. N., and Prescott, T. J. (2003). A basal ganglia inspired model of action selection evaluated in a robotic survival task. *Journal of Integrative Neuroscience*, 2(2):179–200.

Girard, B., Filliat, D., Meyer, J.-A., Berthoz, A., and Guillot, A. (2005). Integration of navigation and action selection in a computational model of cortico-basal ganglia-thalamocortical loops. *Adaptive Behavior: Special Issue on Artificial Rodents*, 13(2):115–130.

Gurney, K., Humphries, M., Wood, R., Prescott, T., and Redgrave, P. (2004). Testing computational hypotheses of brain systems function: a case study with the basal ganglia. *Network: Computation in Neural Systems*, 15:263–290.

Gurney, K., Prescott, T. J., and Redgrave, P. (2001a). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, 84:401–410.

Gurney, K., Prescott, T. J., and Redgrave, P. (2001b). A computational model of action selection in the basal ganglia. II. Analysis and simulation of behaviour. *Biological Cybernetics*, 84:411–423.

Humphries, M. D. and Gurney, K. N. (2002). The role of intra-thalamic and thalamocortical circuits in action selection. *Network: Computation in Neural Systems*, 13:131–156.

Juraska, J., Wilson, C., and Groves, P. (1977). The substancia nigra of the rat: a golgi study. *Journal of Comparative Neurology*, 172:585–600.

Kimura, A. and Graybiel, A., editors (1995). *Functions of the Cortico-Basal Ganglia Loop*. Springer, Tokyo/New York.

Kita, H., Tokuno, H., and Nambu, A. (1999). Monkey globus pallidus external segment neurons projecting to the neostriatum. *Neuroreport*, 10(7):1476–1472.

Krotopov, J. and Etlinger, S. (1999). Selection of actions in the basal ganglia thalamocortical circuits: Review and model. *International Journal of Psychophysiology*, 31:197–217.

Lohmiller, W. and Slotine, J. (1998). Contraction analysis for nonlinear systems. *Automatica*, 34(6):683–696.

Lohmiller, W. and Slotine, J. (2000). Nonlinear process control using contraction analysis. *American Institute of Chemical Engineers Journal*, 46(3):588–596.

Middleton, F. A. and Strick, P. L. (1994). Anatomical evidence for cerebellar and basal ganglia involvement in higher cognitive function. *Science*, 266:458–461.

Mink, J. W. (1996). The basal ganglia: Focused selection and inhibition of competing motor programs. *Progress in Neurobiology*, 50(4):381–425.

Montes-Gonzalez, F., Prescott, T. J., Gurney, K. N., Humphries, M., and Redgrave, P. (2000). An embodied model of action selection mechanisms in the vertebrate brain. In Meyer, J.-A., Berthoz, A., Floreano, D., Roitblat, H., and Wilson, S. W., editors, *From Animals to Animats 6*, volume 1, pages 157–166. The MIT Press, Cambridge, MA.

Parent, A., Sato, F., Wu, Y., Gauthier, J., Lévesque, M., and Parent, M. (2000). Organization of the basal ganglia: the importance of the axonal collateralization. *Trends in Neuroscience*, 23(10):S20–S27.

Park, M., Falls, W., and Kitai, S. (1982). An intracellular hrp study of rat globus pallidus. i. responses and light microscopic analysis. *Journal of Comparative Neurology*, 211:284–294.

Redgrave, P., Prescott, T. J., and Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89(4):1009–1023.

Sato, F., Lavallee, P., Lévesque, M., and Parent, A. (2000). Single-axon tracing study of neurons of the external segment of the globus pallidus in primates. *Journal of Comparative Neurology*, 417:17–31.

Slotine, J. (2003). Modular stability tools for distibuted computation and control. *Journal of Adaptive Control and Signal Processing*, 17(6):397–416.

Staines, W., Atmadja, S., and Fibiger, H. (1981). Demonstration of a pallidostriatal pathway by retrograde transport of hrp-labelled lectin. *Brain Research*, 206:446–450.

Wilson, C. (1993). The generation of natural firing patterns in neostriatal neurons. In Arbruthnott, G. and Emson, P., editors, *Chemical signalling in the basal ganglia*, volume 99 of *Progress in Brain Research*, pages 277–297. Elsevier, Oxford.

Wu, Y., Richard, S., and Parent, A. (2000). The organization of the striatal output system: a single-cell juxtacellular labeling study in the rat. *Neuroscience Research*, 38:49–62.

The basal ganglia as the selection mechanism in a cognitive task*

Tom Stafford & Kevin Gurney
Department of Psychology, University of Sheffield
Western Bank, Sheffield, S10 2TP, UK
t.stafford@shef.ac.uk

Abstract

This paper builds on our existing, biologically constrained, model of the basal ganglia, which was originally constructed under the premise that these subcortical structures perform action selection. Here we show how this same model, when used in conjunction with a connectionist model of processing in the Stroop task, can provide an improved account of human performance on that task. Our model accounts for a wide variety of phenomenon, and provides a framework for connecting Stroop processing with the neuroanatomical basis of action selection. This work validates modelling the basal ganglia as the vertebrate solution to the action selection problem and demonstrates the importance of action selection issues to understanding performance on cognitive tasks. Proposals are made concerning the desirable properties a selection mechanism must possess.

1 The basal ganglia as a vertebrate solution to the selection problem

> *'A selection problem arises whenever two or more competing systems seek simultaneous access to a restricted system.'* [Redgrave *et al*., 1999]

It has been proposed that the basal ganglia is the vertebrate solution to the selection problem [Redgrave *et al*., 1999]. In other words, that it resolves the competition between different neural command centres requesting behavioural control. This need for selection is most clear, and has been most thoroughly empirically explored, in terms of motor expression, but it is expected that similar functional architecture, and comparable functional requirements, underly selection in different domains.

The basal ganglia has external and internal connectivity that makes it suitable for performing the role of a selection mechanism. It receives inputs from virtually the entire cerebral cortex, limbic system structures such as the hippocampus and the amygdala, and, notably, the anterior cingulate cortex [Masterman and Cummings, 1997; Redgrave *et al*., 1999].

*This work was supported in part by EPSRC grant EP/C516303/1 to Kevin Gurney.

The main input nucleus of the basal ganglia is the striatum, which provides the first processing of incoming signals from cortex. The projection neurons of the striatum (medium spiny neurons) are by default quiescent (in a down-state), and do not do not respond to low levels of input. Only after a substantial and coordinated excitatory input do they move to an up-state, in which they produce significant output which may subsequently be affected by smaller changes in input [Wilson, 1995].

Outputs from the basal ganglia project back to the cortex, via the thalamus, and to premotor areas of the brainstem. The output nucleus of the Basal ganglia is the globus pallidus, internal segment (GPi). Neurons here are are tonically active, inhibiting their target structures from enacting behaviour. Actions are enabled by the selective release of that inhibition. It is posited that signals from cortex indicate the 'salience' — i.e. the importance and urgency — of possible actions to the basal ganglia [Redgrave *et al*., 1999]. Sufficiently large saliences result in the selective disinhibition of channels associated with that action, and thus the release of the action [Chevalier and Deniau, 1990].

1.1 Modelling confirms that the basal ganglia can perform action selection

We have constructed a neuronal network model of the basal ganglia, constrained by the known anatomy and physiology, and based on the selection hypothesis [Gurney *et al*., 2001a; 2001b]. Analysis and simulation of this model [Gurney *et al*., 2001b] shows that the basal ganglia display the properties of a good selection mechanism [Redgrave *et al*., 1999]: the highest salience rapidly promotes appropriate channel selection; once selection has been made competitors do not distort that selection; however, significant changes in the salience inputs result in rapid and clean channel switching.

Embedding this model into its anatomical context provided by cortex and thalamic circuits, improves the selection behaviour and gives a more complete understanding of the functional role the different nuclei involved may be playing [Humphries and Gurney, 2002]. Further, using these models in robot controllers shows that the selection behaviour is of sufficient efficiency and sophistication to be behaviourally adequate in realistic environments [Girard *et al*., 2003; Montes-Gonzalez *et al*., 2000].

The simulation work presented below uses the basal gan-

glia model exactly as presented elsewhere [Gurney *et al*., 2001b; Humphries and Gurney, 2002]. We focus on the benefits of using this biologically plausible mechanism, which has been demonstrated to possess ethologically realistic selection properties. The internal structure of the basal ganglia model is only discussed as far as is necessary to illustrate *why* it works as it does in the context of the current work.

1.2 Using the basal ganglia model in a cognitive task

This paper is concerned with a different extension of the model— into the domain of cognitive selection and performance, as measured by reaction times. In particular we consider a celebrated cognitive task that involves a selection conflict — the Stroop Task [Stroop, 1935] — and take as our starting point the most successful computational model of performance on this task to date [Cohen *et al*., 1990] which we extend by integration with our existing model of basal ganglia function [Gurney *et al*., 2001a; 2001b; Humphries and Gurney, 2002].

The purposes of this extension are threefold. Firstly, it allows an additional test of the basal ganglia model of action selection. The model was constructed using the known functional neuroanatomy and guided by the selection hypothesis. It was not explicitly designed to simulate reaction times, nor was it constrained by human cognitive performance. However, any action selection mechanism should also be able to act as a response selection mechanism in cognitive tasks. Therefore the performance of the model in this domain is a good test of its validity. Secondly, some aspects of human Stroop performance remain inadequately addresses by existing models, leaving the possibility open that a model containing new elements may improve the possible account and shed light on why previous models have not been so successful. Additionally, making connection to the possible underlying neurobiology enriches the account possible of Stroop processing. In particular, we anticipate that features of the basal ganglia model such as allowing arbitrary numbers of inputs and making provision for dopaminergic modulation of signal processing will provide opportunities for future experimental and modelling investigations. Thirdly, integrating cognitive and systems-neuroscience models sheds light on issues of selection from both levels of analysis. We will attempt to use our combined model to derive some general constraints on models of selection.

2 Modelling the Stroop Task

2.1 The Stroop Task

J. Ridley Stroop's famous task [Stroop, 1935] involves presenting words written in coloured inks. Participants must name the colour of the ink while trying to ignore the word, which can spell out the name of a colour. When the word-name is in contradiction to the ink-colour the task becomes effortful, slowed and error-prone. This is the interference effect, traditionally measured as the difference in reaction time (RT) or errors between the control condition (when the word-aspect of the stimuli is nominally neutral with respect to colour) and the conflict condition (when the word-aspect of the stimulus contradicts the color). There is a corresponding facilitation effect; when the word and the colour aspect match (the congruent condition) there is a speeding relative to the control condition. These two effects are asymmetrical; facilitation is typically far smaller than interference. The converse task — reading the word while ignoring the ink-colour — can also be assessed. Word-reading is faster than colour-naming, and is not affected by the colour-aspect of the stimulus (there is no interference or facilitation).

Traditionally the Stroop task has been discussed in terms of a conflict between automatic and controlled processes [MacLeod, 1991], and much progress has been made in using variations of the Stroop task to adumbrate the nature of 'automatic' processing [Besner and Stolz, 1999; Besner *et al*., 1997; Dishon-Berkovits and Algom, 2000; Durgin, 2000]. But it is also apparent that the Stroop task involves a selection conflict and provides a thoroughly explored experimental framework for investigating cognitive aspects of selection.

2.2 A Model of the Stroop task

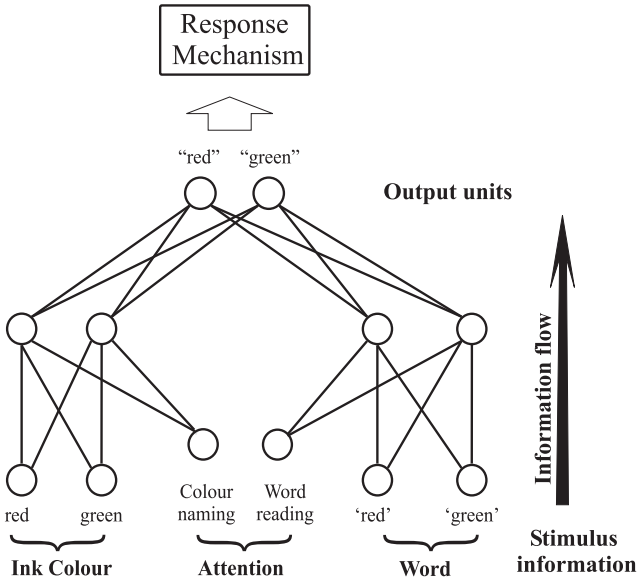


Figure 1: Architecture of the Cohen model

By far the most successful quantitative model of Stroop processing is that of Cohen *et al* [1990]. This simple connectionist model (hereafter 'The Cohen model') involves the translation of a localist input representation into a response representation, via a feed-forward two-layer network trained with backpropagation (Figure 1). The main features of this network are:

1. Differential training of the network: responses to word inputs are trained at ten times the frequency of responses to colour inputs. This results in a stronger weighting of signals representing this aspect of the stimulus.
2. Attentional sensitisation: the network implements attention as an additional input which off-sets a bias (in effect

a default inhibition) on all hidden units. This interacts with the sigmoidal output function of the units so that moderately sized signals do not result in a commensurate increase in output unless presented in combination with attentional input. Signals in the word-processing pathway, however, are large enough to partially overcome the default inhibition without the aid of attentional input.

3. Reaction times generated a by response mechanism that works on evidence accumulation: the output units of the network are taken to indicate, at each time point, the evidence favouring each response. This evidence is compared and accumulated until the total crosses a threshold – when a response is said to have been made. It is this feature feature of the model which is the concern of the current work.

Selection in Cohen *et al*'s (1990) model of the Stroop Task

The response mechanism of Cohen *et al*'s [1990] model is ignored in textbook treatments of the model [Ellis and Humphreys, 1999; Sharkey and Sharkey, 1995] and even overlooked in Cohen *et al*'s own analysis of the function of the model [Cohen *et al*., 1990]. This reflects, we argue, a regrettable, but not untypical, neglect of the action selection problem in psychology. Reinforcing this view, we have recently, shown that, contrary to the original account of Cohen *et al*., it is the response mechanism, not the neuronal transfer function, which generates the important differences in reaction times between conditions [Stafford and Gurney, 2004; Stafford, 2003], and it is the response mechanism which explains the asymmetry in the magnitudes of the interference and facilitation effects in the Cohen model (a matter about which there has been some debate [MacLeod and MacDonald, 2000]). The response mechanism of the model is isomorphic to the diffusion model [Ratcliff, 1978; Ratcliff *et al*., 1999], which has been shown to be an analytically tractable form of several connectionist models of decision, and an optimal decision algorithm for a two-choice decision situation [Bogacz *et al*., submitted] where either desired accuracy or time-to-decision is specified (obviously these two mutually constrain each other). Further, potential neurobiological correspondences to the evidence accumulation processes of the diffusion model have been identified [Gold and Shadlen, 2000]. Thus our investigation of evidence accumulation as a mechanism of selection in this specific model may carry important lessons for theories of selection in general.

The response mechanism is also responsible for a major mismatch between model performance and human performance. A plausible alternative theory of Stroop processing — and of automatic processing in general — is that more automatic processes are simply faster. This theory would suggest that Stroop interference is due to the response evoked by the word aspect of the stimulus arriving at some response bottleneck earlier, creating slower selection of the opposite response when it arrives there. The experimental refutation of this theory involves presenting a coloured-ink patch next to a colour-word. If presented simultaneously the normal Stroop effect is found, but the spatial separation allows the asynchronous presentation of the colour and the word; a stimulus onset asynchrony (SOA) paradigm. If the colour appears sufficiently before the word then, according to the simple 'horse-race' theory, naming of the word should suffer interference from the colour information (a 'reverse Stroop effect'). This is not what happens experimentally [Glaser and Glaser, 1982]. For word-naming, no amount of head-start for colour-information is sufficient to create interference. For colour-naming, the appearance of the word at any point up to 300 ms after the appearance of the colour (close to the asymptotic limit for reaction times) causes interference. Additionally, the appearance of the word before the colour always causes interference, however long the subject is given to accommodate to the presence of the word.

The Cohen *et al* model can simulate limited features of the Stroop SOA paradigm. However, if the model is tested beyond the range presented in the original paper, serious flaws are revealed. Trends in the simulation data which can be seen over the original range of SOA values continue at longer SOAs, as the to-be-ignored dimension of the stimulus is presented increasingly before the to-be-responded to dimension (see Figure 2). By convention SOAs which involve the to-be-ignored dimension being presented first are labelled negative. Thus, for colour naming, in the conflict condition, the model response time increases as the SOA gets more negative until eventually the word-aspect is presented early enough to force an incorrect response. In Figure 2 this is represented by the peak in the line showing the colour-naming conflict condition reaction times. RTs start to decrease with increasing negative SOA because the model is more and more quickly selecting the wrong response. If the word is congruent to the colour information then there is comparable interference, but this reveals itself as a speeding of the correct response; this dynamic continues until, ultimately, the model responds before the colour information has even been presented. In Figure 2 this shown by the point at which the line representing RTs for colour-naming congruent condition crosses the dotted line representing zero on the RT axis. For the same fundamental reasons, in the word-naming task the conflict and congruent conditions diverge in the same way (albeit over a longer time span). Thus, the model behaves in accordance with the experimentally disproved horse-race model: presenting colour information ahead of word information creates a reverse Stroop effect — colour information interferes with word-reading.

The reason for these failures may be traced to the evidence accumulation response mechanism. Because the model, like all connectionist models, works on graded signals there is always some signal change due to the to-be-ignored, even if this is very small due to the attentional inhibition. In the case of the colour-naming task, it is integral to the model's function that some influence of the word-aspect of the stimulus survives attentional selection and comes to influence the response stage. Without this feature the basic effect of Stroop interference would not be present. However, in SOA conditions, this influence of the to-be-ignored aspect may accumulate indefinitely. This affects selection time to an extent proportional to the time it is presented multiplied by the strength of evidence conveyed. So arbitrarily small amounts of evidence can provoke erroneous selection if presented for long

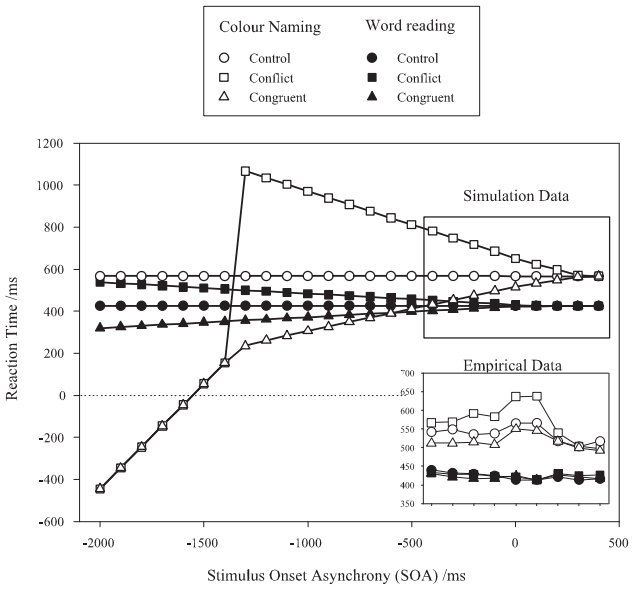


Figure 2: The SOA simulation of the original Cohen model. The empirical data is shown inset. The simulation data corresponds roughly to the empirical data over the range originally reported (-400 to +400 ms) but beyond that diverges.

enough, or they can massively slow correct selection (because accumulated evidence for the opposite response must be overcome).

Adding a more biologically realistic response mechanism — based on the basal ganglia — overcomes these deficiencies and considerably extends the model's explanatory power.

3 The Basal Ganglia model as a response mechanism for a cognitive task

The neural network component of Cohen *et al*'s model performs what is normally thought of as the cognitive elements of the task: stimulus–response translation, attentional control and learning. Only one minor change was required to this 'front-end' to make is compatible with using the basal ganglia model as the response mechanism. The output units of the Cohen model originally had resting values of 0.5. This was changed to 0.1, to make the output signals interpretable by the basal ganglia model as indicative of the salience of the corresponding response [1].

In all other respects the combined model is exactly as published by Cohen *et al* [1990], except with the basal ganglia model [Gurney *et al*., 2001b; Humphries and Gurney, 2002] replacing evidence accumulation as the method of final response selection.

[1]For consistency this entails changes in the initial weights the networks is given before training, but these are not discussed here as there is no substantive effect on the simulation results; as should be expected from a good model the principle findings are robust under parametric variation, and this aspect of the model is an implementational detail which is irrelevant to overall behaviour of the model. For details see Stafford [2003].

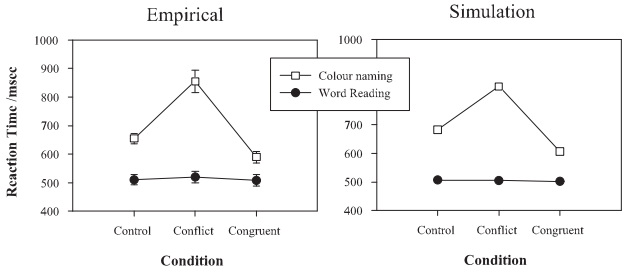


Figure 3: Empirical and simulation reaction times in the basic Stroop conditions for word-reading and colour-naming tasks. Empirical data is from Dunbar & MacLeod [1984], for which standard error bars are shown

4 The Simulation Results

4.1 Matching and Improving on the performance of Cohen's Model

We tested the Cohen connectionist front-end with the basal ganglia model as the response mechanism (hereafter 'The Model') on the first three simulations presented by Cohen *et al* [1990]. The model simulates the basic Stroop task (simulation 1), matching the empirical data as well as the original Cohen model does (Figure 3).

The ability to realistically model learning phenomena is a key benefit of connectionist models. The model mimics the power-law function of learning (Figure 4), just as the original Cohen model does. This demonstrates that the learning dynamic captured by the connectionist front-end is not interfered with by the use of the basal ganglia response mechanism; graded changes in the signals from the front end are converted into appropriately graded changes in reaction times.

The SOA task (Simulation 2) shows up the superiority of the basal ganglia as a response mechanism over the original response mechanism. As discussed, over long negative SOAs the Cohen response mechanism makes wrong selections, due to the small but significant influence of the distracting stimulus dimension. The input units of the basal ganglia model filter out small salience inputs (as discussed section 1). This creates a minimal salience threshold, below which inputs are ignored. Thus, using the basal ganglia response mechanism, the model makes the correct selection at all SOA values. Furthermore, the distracting influence of the to-be-ignored aspect of the stimulus is limited. This is reflected in the stabilisation of reaction times at SOAs below –400 ms (see Figure 5).

5 General Discussion

5.1 Strengths of the Model

This work validates our model against the basic Stroop phenomena. Use of the basal ganglia model as the response mechanism improves the fit that can be made to the empirical data and highlights necessary features response mecha-

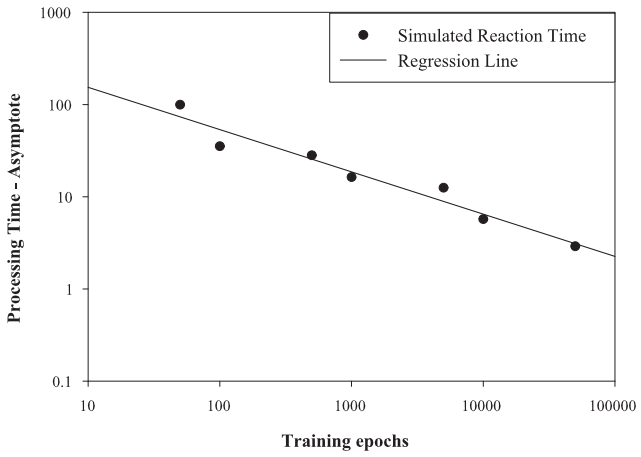


Figure 4: The model conforms to the power law of practice [Logan, 1988]. Both axis use a log scale. Simulation results are shown as dots. The simple regression for the data is shown as a straight line and follows the form log_{10}(Processing Time) $= 2.65 - 0.46 \times log_{10}$(Epochs). $R^2 = 0.948$.

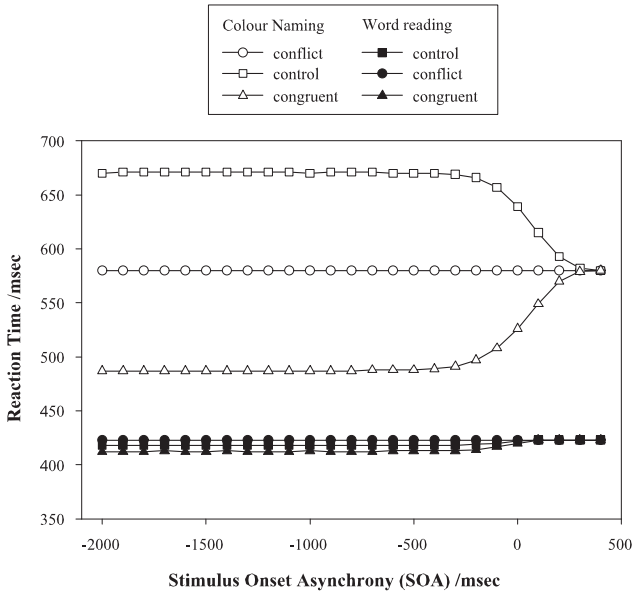


Figure 5: Model SOA data.

nisms should contain, the lack of which was overlooked in the previous account. Use of the basal ganglia model also extends the account of Stroop processing to connect with the neurobiology of selection. The basal ganglia model includes anatomical specific pathways and an account of the dopamine system. This allows future tests of the model against various pathologies, such as schizophrenia.

A better account of the data is one benefit of this model. There is also a theoretical purity to testing models by utilising them in new areas that they were not developed with in mind. It is testament to the basal ganglia model's value as a general model of selection that it deals appropriately with signals provided by a connections model of a cognitive task.

5.2 Why Does The Model Work?

The model captures the basic Stroop (Figure 3) and learning (Figure 4) phenomena because, for moderately sized saliences, selection time is based on the relative difference between the to-be-selected salience and the competing salience (if any). It is with small saliences, and when dealing with successive rather than simultaneous inputs, that the basal ganglia model shows its superiority as a selection mechanism. Both of these cases are revealed by comparison of the SOA simulations (Figures 2 and 5).

The failure of the Cohen model on the SOA simulations is because of a model feature which is neither trivial nor irrelevant. The existence empirically of the basic Stroop interference effect demonstrates that response activation from the to-be-ignored word aspect of the stimulus must break through any initial attentional inhibition. Arriving at the response mechanism before the response activation of the colour aspect, this activity is enough, in Cohen's model, to cause selection. The erroneous selection produced at long SOAs shows that a response mechanism must not make selections based on inconsequentially low inputs. Our basal ganglia model avoids this by having a minimum salience threshold, below which no action is selected.

This minimal threshold was included in the basal ganglia model because of the neurobiology of medium spiny neurons in the striatum – the main input nucleus of the basal ganglia. These neurons possess upstate / downstate functionality, which means that they only start to release action potentials if their input is above a certain threshold. This feature has the effect of filtering out noise in the inputs which is below threshold. The Cohen model evidence accumulation mechanism has no such minimal threshold, and no decay of accumulated evidence, and because of this it always makes a selection if left for long enough. By extension, the diffusion model, the general form of the evidence accumulation mechanism used, contains no capacity for not making a selection. This is a serious flaw. It means that evidence accumulation and the diffusion model alone cannot provide a full account of action selection.

In the basal ganglia model, competition on other channels, even if below selection level, can affect selection time. This priming, whether positive or negative, occurs because activity on other channels alters the resting level of output signals in GPi, and thereby affects the time it takes for outputs to drop to the point whereby selection occurs. The amount of this priming is limited because the model uses units with an output range restricted between 0 and 1. Compare this with the Cohen response mechanism, and by extension the diffusion model, which, with no constraints on where the selection threshold is set, contains the capacity to retain infinitely large values and thus can generate arbitrarily large amount of interference (as seen in the SOA simulations, Figure 2). This benefit of the basal ganglia model demonstrates the value of considering the mechanisms of action selection within a (neural) signal processing context.

5.3 What properties must a selection mechanism possess?

At a minimum these issues indicate that the context within which the diffusion model of selection is used cannot be ignored or assumed. The simulation of the SOA paradigm highlights two properties which the basal ganglia as a selection mechanism brings to the combined model to improve the account of the data. Together both of these features mean that not only is the wrong response not selected, but also the right response is selected efficiently. This is an example of the 'clean switching' property which has been identified as a necessary feature of any selection mechanism [Redgrave *et al*., 1999].

The first feature is that the basal ganglia model limits the maximum possible influence on selection of concurrently or consecutively active competing inputs. So, in the SOA paradigm with negative SOAs the interference on reaction time does not get progressively longer with increasing SOA, but instead levels off – there is a maximum amount of interference that a distracting stimulus can produce on reaction times. This benefit is due to the wider context of adaptive control that the basal ganglia model arose from. A response mechanism needs to work in real-time, continuously, dealing with the successive selection of actions and interruption of old actions by new. Because the BG model is designed to operate continuously it has equilibrium final states, in which no action is selected. All patterns of input, if unchanging, eventually produce unchanging output states (although such a situation is unlikely to arise). For some patterns of input, the final output state indicates that no action is selected. The evidence accumulation response mechanism, on the other hand, has only one type of final state – that of selecting an action – and it continuously moves towards this state. The existence of equilibrium final states allows the successive switching between actions, without those actions interfering more with the selection of new actions the longer they have been selected.

The second necessary feature is that the basal ganglia will not make selections based on arbitrarily low inputs. This is because, due to the physiological properties of the medium spiny neurons in the striatum, the input nucleus of the basal ganglia, the model has built into it a minimum input threshold below which signals are ignored. Without such an input threshold, any level of input will cause the evidence accumulation counter to inextricably increase towards the selection threshold. A minimum input threshold is not the only way of preventing this kind of erroneous selection. Usher & McClelland [2001], in their model of perceptual choice, present an alternative strategy to a minimal input threshold, but one which has the same functional role. They argue that models of perceptual choice – they discuss the same kind of choice algorithms that are the basis for the Cohen *et al* [1990] response mechanism [Luce, 1986] – require the addition of activation decay on the choice representations. A decay mechanism can fulfill the same role as a minimal input threshold, since for situations where input is less than the decay that input is effectively filtered out. Another way of solving this erroneous selection problem might be to send a no-go signal which prevents selection until appropriate. This would not be feasible with a evidence accumulation model of selection, but it would be feasible with the basal ganglia response mechanism because it does not allow previous signal values to carry potentially unlimited weight when selecting new actions (i.e. clean switching, as discussed above). In the SOA task a no-go signal could be provided by the front-end to the basal ganglia on a third channel. This no-go signal, by being itself selected, can prevent selection until the relevant stimulus dimension has appeared. Although possible, this type of solution is perhaps not theoretically desirable because it relegates the problem of selection to another part of the system and hence begs the question of how correct selection is achieved.

A third possible way of accounting for the basic Stroop effect but avoiding erroneous selection in the SOA conditions is to include in the model a kind of reactive attentional inhibition, which suppresses activity based on the to-be-ignored dimension but only after it has occurred. Just such a stimulus-evoked inhibition mechanism is the focus of the cognitive control hypothesis of Botvinick *et al* [2001]. Initial investigations suggest that this mechanism, because it reduces interference from the to-be-ignored dimension of the Stroop stimulus but not until that interference has first arisen, would allow the accurate modelling of the course of interference in the SOA paradigm [Stafford, 2003]. Future modelling work may suggest ways in which these ways of limiting interference and preventing selection based on arbitrarily small values can be experimentally distinguished.

References

[Besner and Stolz, 1999] D. Besner and J. Stolz. Context dependency in stroop's paradigm: When are words treated as nonlinguistic objects? *Canadian Journal of Experimental Psychology*, 53(4):374–380, 1999.

[Besner *et al*., 1997] D. Besner, J. A. Stolz, and C. Boutilier. The stroop effect and the myth of automaticity. *Psychonomic Bulletin & Review*, 4(2):221–225, 1997.

[Bogacz *et al*., submitted] R. Bogacz, E. Brown, J. Moehlis, P. Hu, P. Holmes, and J.D. Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced choice tasks. submitted.

[Botvinick *et al*., 2001] M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, and J. D. Cohen. Conflict monitoring and cognitive control. *Psychological Review*, 108(3):624–652, 2001.

[Chevalier and Deniau, 1990] G. Chevalier and J.M. Deniau. Disinhibition as a basic process in the expression of striatal functions. *Trends in Neurosciences*., 13:277–281, 1990.

[Cohen *et al*., 1990] J. D. Cohen, K. Dunbar, and J. L. McClelland. On the control of automatic processes - a parallel distributed-processing account of the stroop effect. *Psychological Review*, 97(3):332–361, 1990.

[Dishon-Berkovits and Algom, 2000] M. Dishon-Berkovits and D. Algom. The stroop effect: It is not the robust phenomenon that you have thought it to be. *Memory & Cognition*, 28(8):1437–1449, 2000.

[Dunbar and MacLeod, 1984] K. Dunbar and C. M. MacLeod. A horse race of a different color - stroop interference patterns with transformed words. *Journal of Experimental Psychology - Human Perception and Performance*, 10(5):623–639, 1984.

[Durgin, 2000] F. H. Durgin. The reverse stroop effect. *Psychonomic Bulletin & Review*, 7(1):121–125, 2000.

[Ellis and Humphreys, 1999] R. Ellis and G. Humphreys. *Connectionist Psychology : a text with readings*. Psychology Press Ltd, Hove, UK, 1999.

[Girard *et al*., 2003] B. Girard, V. Cuzin, A. Guillot, K. N. Gurney, and T. J. Prescott. A basal ganglia inspired model of action selection evaluated in a robotic survival task. *Journal of Integrative Neuroscience*, 2(2):179–200, 2003. Journal version of the SAB paper.

[Glaser and Glaser, 1982] M. O. Glaser and W. R. Glaser. Time course analysis of the stroop phenomenon. *Journal of Experimental Psychology-Human Perception and Performance*, 8(6):875–894, 1982.

[Gold and Shadlen, 2000] J. I. Gold and M. N. Shadlen. Representation of a perceptual decision in developing oculomotor commands. *Nature*, 404(6776):390–394, 2000.

[Gurney *et al*., 2001a] K. Gurney, T.J. Prescott, and P. Redgrave. A computational model of action selection in the basal ganglia i: A new functional anatomy. *Biological cybernetics*, 85(6):401–410, 2001.

[Gurney *et al*., 2001b] K. Gurney, T.J. Prescott, and P. Redgrave. A computational model of action selection in the basal ganglia: Ii: Analysis and simulation of behaviour. *Biological cybernetics*, 85(6):411–423, 2001.

[Humphries and Gurney, 2002] M. D. Humphries and K. N. Gurney. The role of intra-thalamic and thalamocortical circuits in action selection. *Network-Computation in Neural Systems*, 13(1):131–156, 2002.

[Logan, 1988] G.D. Logan. Toward an instance theory of automatization. *Psychological Review*, 95(4):492–527, 1988.

[Luce, 1986] R.D. Luce. *Response Times: Their Role in Inferring Elementary Mental Organisation*. Oxford Psychology Series. Clarendon Press, New York, 1986.

[MacLeod and MacDonald, 2000] C.M. MacLeod and P.A. MacDonald. Interdimensional interference in the stroop effect: uncovering the cognitive and neural anatomy of attention. *Trends in Cognitive Sciences*, 4(10):383–391, 2000.

[MacLeod, 1991] C. M. MacLeod. Half a century of research on the stroop effect - an integrative review. *Psychological Bulletin*, 109(2):163–203, 1991.

[Masterman and Cummings, 1997] D. L. Masterman and J. L. Cummings. Frontal-subcortical circuits: The anatomic basis of executive, social and motivated behaviors. *Journal of Psychopharmacology*, 11(2):107–114, 1997.

[Montes-Gonzalez *et al*., 2000] F. M. Montes-Gonzalez, T. J. Prescott, K. Gurney, and P. Redgrave. The robot basal ganglia: Control of robot action selection by an embodied model of the mammalian basal ganglia. *European Journal of Neuroscience*, 12:134–134, 2000.

[Ratcliff *et al*., 1999] R. Ratcliff, T. Van Zandt, and G. McKoon. Connectionist and diffusion models of reaction time. *Psychological Review*, 106(2):261–300, 1999.

[Ratcliff, 1978] R. Ratcliff. A theory of memory retrieval. *Psychological Review*, 85:59–108, 1978.

[Redgrave *et al*., 1999] P. Redgrave, T. J. Prescott, and K. Gurney. The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience*, 89(4):1009–1023, 1999.

[Sharkey and Sharkey, 1995] A.J.C. Sharkey and N.E. Sharkey. Cognitive modeling: psychology and connectionism. In M.A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 200–203. The MIT Press, Cambridge, MA., 1995.

[Stafford and Gurney, 2004] T. Stafford and K.N. Gurney. The role of response mechanisms in determining reaction time performance: Pieron's law revisited. *Psychonomic Bulletin & Review*, 11(6):975–987, 2004.

[Stafford, 2003] T. Stafford. *Integrating Psychological and Neuroscientific Constraints in Models of Stroop Processing and Action Selection*. Phd thesis, University of Sheffield. Available at http://www.shef.ac.uk/ ∼abrg, 2003.

[Stroop, 1935] J.R. Stroop. Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, 18:643–662, 1935.

[Usher and McClelland, 2001] M. Usher and J. L. McClelland. The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108(3):550–592, 2001.

[Wilson, 1995] C. Wilson. The contribution of cortical neurons to the firing pattern of striatal spiny neurons. In J. Houk, J. Davis, and D. Beiser, editors, *Models of information processing in the basal ganglia*, pages 29–50. MIT Press, Cambridge, MA., 1995.

Action Selection in Subcortical Loops through the Basal Ganglia

JC Houk, D Fraser, A Fishbach, SA Roy, LS Simo, C Bastianen,
D Fansler-Wald, LE Miller, PJ Reber, M Botvinick
Northwestern University Institute for Neuroscience and the
Department of Physiology at the Medical School, Chicago,
University of Pennsylvania Department of Psychiatry, Philadelphia
j-houk@northwestern.edu, mmb@mail.med.upenn.edu

Abstract

Subcortical loops through the basal ganglia and cerebellum form computationally powerful distributed processing modules (DPMs). This paper relates the computational features of a DPM's loop through the basal ganglia to experimental results for two kinds of natural action selection. First, functional imaging during a serial order recall task was used to study human brain activity during the selection of sequential actions from working memory. Second, microelectrode recordings from monkeys trained in a step-tracking task were used to study the natural selection of corrective submovements. Our DPM-based model assisted in the interpretation of puzzling data from both of these experiments. We come to posit that the many loops through the basal ganglia each regulate the embodiment of pattern formation in a given area of cerebral cortex. This operation serves to instantiate different kinds of action (or thought) mediated by different areas of cerebral cortex.

1. DPM-based Model

The higher order circuitry of the brain is comprised of a large-scale network of cerebral cortical areas that are individually regulated by loops through subcortical structures, particularly through the basal ganglia and the cerebellum [Houk and Wise 1995; Kelly and Strick 2003, 2004]. These subcortical loops form distributed processing modules (DPMs) that have powerful computational architectures (Figure 1) [Houk 2005]. The final outcome of all of the computations in a given DPM is a spatiotemporal pattern of activity in the module's output vector, representing the activity in its set of cortical output neurons. This allows a given DPM to participate in the computations taking place in other areas of cerebral cortex, or in the brainstem or spinal cord.

The loop through the basal ganglia is thought to regulate the selection and/or initiation of pattern formation [Gurney et al. 2001; Houk & Wise 1995; Houk 2001; Redgrave et al 1999]. The term Embodiment is used in Figure 1 [Houk 2005] to capture both possibilities, i.e. either selection or initiation, the former occurring when disinhibition allows other cortical inputs to initiate and the latter when the selection is strong and does its own initiation. Embodiment is critically dependent on the refined, neuromodulated pattern classification operations that take place in the input layer of the basal ganglia, the striatum [Gruber et al. 2003]. According to most contemporary models, bursts of striatal spiny neurons, via the direct pathway through the basal ganglia, disinhibit their targets in thalamus, allowing thalamo-cortical loops to embody patterns of activity that represent a ballpark estimate of an action, or a thought. In contrast, via the indirect pathways through the basal ganglia, bursts of striatal spiny neurons depress their targets in thalamus, inhibiting the embodiment of patterns that would represent poor choices in action selection.

Once a tentative pattern has been selected and initiated through the operation of the loops through the basal ganglia, the loops through the cerebellum amplify and sculpt that pattern into a refined output vector [Houk and Mugnaini 2002]. The amplification step appears to be implemented by the loop through the cerebellar nuclei. Regenerative positive feedback in this loop amplifies the output's intensity, duration and spatial extent. The restrainment of this amplification process and, more importantly, sculpting it into an accurate representation of an action (or thought) is implemented by the loop through the cerebellar cortex. The cerebellar cortex is considered to be an exceptional neuronal architecture for learning difficult computations [Raymond et al. 1996; Houk & Mugnaini, 2002] and so is well suited to this task.

2. Serial Order Recall

Tasks in which lists of items are presented, after which the subject is required to recall the items in the same order in which they were presented, require serial order processing and sequential action selection. Here we introduce a task dubbed Replicate without intending to identify novel behavioral phenomena. Instead we aspire to establish a task paradigm that elicits many standard patterns of serial recall behavior, but which also can be conveniently applied across research modalities and, in particular, across species. Benchmark properties of serial order recall include: 1) A graded decline in recall accuracy with sequence length, 2) transposition gradients reflecting a tendency for items to be recalled at serial positions near to their original positions, 3) item similarity effects including a) a tendency for items to be recalled near the item where they originally appeared, b) a tendency for sequences of similar items to be recalled

less accurately than sequences of less similar items [Botvinick and Plaut 2005].

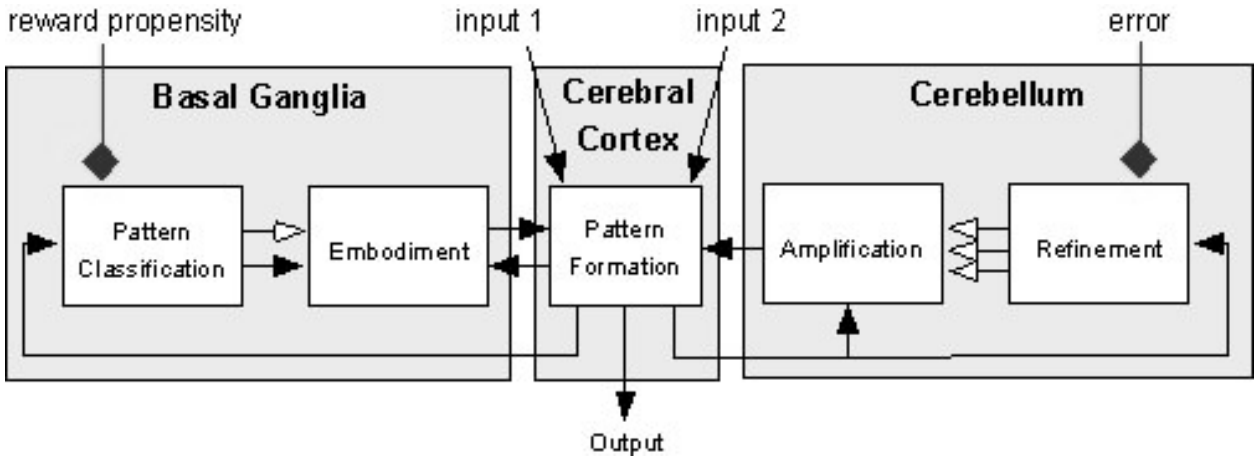


Figure 1: The abstract signal processing operations posited for each DPM. Net excitatory pathways are shown with closed arrows, net inhibitory pathways are shown with open arrows, and the diamonds signify neuromodulatory and training inputs

The Replicate task presents K targets on an NxN grid of squares in a randomized sequence and requires the subjects to remember their positions and serial order over a brief delay. The subjects are then cued to use a joystick to move a cursor to the K positions in the same order in which they were originally presented. The phase of target presentation requires the setting up of a working memory representation, which must be sustained through the delay and then decoded in order to produce correct joystick movements; we thus refer to the three phases of the task as the encoding, maintenance and decoding phases. We also employ a control task, referred to as Chase. In Chase, a sequence of location cues appears just as in Replicate, but subjects use the joystick to track these cues immediately as they appear. Chase involves similar stimulus and response sequences to the Replicate task, but eliminates the working memory component.

Preliminary behavioral studies with Replicate confirm that the task generates several standard patterns of recall behavior. Thirty-two Replicate trials were performed, eight at each of four sequence lengths (3-6 for half the subjects, 4-7 for the other half). Each trial was initiated by the subject using the joystick to move a cursor into the central tile in a 5x5 grid. A target sequence then appeared, with each target location illuminated for a total of 500 msec. Following a 10 sec delay, the joystick cursor changed color, cuing the subject to reproduce the target sequence, returning to the central tile when finished. A maximum of 3 sec was allotted for identification of each location. Our error analysis suggested that the Replicate task yields the typical visual memory span of 4-5 items, and that errors frequently involve 1) transpositions of items located near to one another in the sequence and/or 2) substitution of a location target with a nearby location in the grid. These results demonstrate that Replicate has several benchmark properties of serial order recall [Botvinick and Plaut 2005].

Functional neuroimaging (fMRI) was used to study BOLD changes during the Replicate task. There were two primary BOLD contrasts. The "execution" contrast was made between the period of sensory guided joystick movements in the Chase task and a rest period. This contrast was designed to show the neural correlates of serial motor execution. The "decoding" contrast was made between the memory guided movement period of the Replicate task and the sensory guided movement period of the Chase task. This contrast was designed to reveal the neural correlates of the decoding process while simultaneously controlling for BOLD activity related to pure motor execution. Whole brain EPI data (24 6 mm slices, TR=2000 ms) were collected from 10 subjects, and a partial-brain scanning protocol focusing on the basal ganglia (12 6mm slices, TR=1000 ms) was used for 9 subjects.

In the participants who provided whole-brain data, reliable decoding activity was observed in right prefrontal cortex, left anterior cingulate, left supplementary motor area, and portions of cerebellum. Activity related to the execution of joystick movements was observed in the contralateral primary motor cortex, contralateral putamen, and ipsilateral cerebellar cortex. The partial-brain imaging protocol provided better sensitivity to changes within the striatum of the basal ganglia.

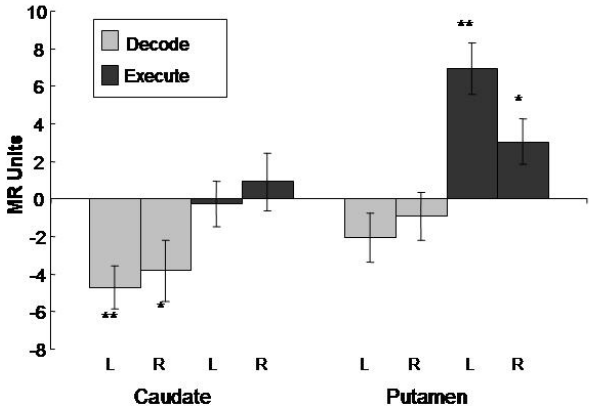


Figure 2: Differential BOLD activity in the right and left head of the caudate and putamen for the decoding (light grey), and execution (dark grey) contrasts. Error bars indicate standard error. Single asterisk (*) indicates a significant difference [$t(8) \geq 2.36$, $p < 0.05$] while double asterisks (**) indicate a highly significant difference [$t(8) \geq 4.16$, $p < 0.01$]. A significant decrease in activity was found in the caudate nucleus for decoding, whereas a significant increase in activity was found in the putamen for execution. Deactivation, representing a statistically significant decrease in blood flow in caudate during decoding, was surprising.

Action selection in the loop through the basal ganglia. Although many authors have suggested that the loop through the basal ganglia plays an important role in action selection, there are diverse views concerning the mechanism by which this might occur. Most authors agree that action selection occurs in the input nucleus of the basal ganglia loop, namely the striatum (but see [Rubchinsky et al. 2003]), comprised of caudate and putamen divisions. The principal neurons of the striatum, the medium spiny neurons, are inhibitory GABAergic projection neurons and emit an elaborate array of collaterals to neighboring spiny neurons before they project to globus pallidus. The collaterals give rise to an inhibitory feedback network in the striatum mediating a competitive pattern classification operation. Collateral inhibition is deemed an effective mechanism for competition by some authors [Plenz 2003] and ineffective by others, the latter believing that feedforward inhibition mediates the pattern classification operation [Tepper et al. 2004]. Beiser and Houk [1998] modeled both mechanisms and found that both worked, but that the inhibitory feedback network worked more effectively than the feedforward network.

What has not been considered to date is the possibility that the inhibitory feedback network relies on presynaptic, as opposed to postsynaptic, inhibition. This is surprising since presynaptic inhibition of cortical input to the striatum has been demonstrated electrophysiologically [Calabresi et al. 1991; Nisenbaum et al. 1993]. Indeed, the operation of a presynaptic mechanism for collateral inhibition could also explain the surprising fMRI BOLD deactivation that we found for the decoding contrast in caudate (Figure 2). Synaptic input is believed to be a strong contributor to BOLD signals (Arbib et al. 2000). Since presynaptic inhibition would decrease synaptic input, that could explain the deactivation for caudate. The activation seen for putamen presumably results from a greater dependence on postsynaptic inhibition. The cause for this difference might relate to phylogeny; by and large, caudate is phylogenetically more recent than putamen.

Model of competitive pattern classification. Presynaptic inhibition should give rise to a computationally powerful mechanism for pattern classification. Beiser and Houk [1998] found that, since the equilibrium potential for postsynaptic GABAergic inhibition, is between the down and up state of spiny neurons, this mechanism for mediating competition between neighboring spiny neurons was quite sensitive to spontaneous membrane potential and to model parameters. It performed better than feedforward inhibition, but it was not optimal. Presynaptic inhibition has no equilibrium potential – it just cuts off the synaptic input regardless of the membrane potential of the spiny neuron.

Fansler-Wald et al [2004] modeled a network of recurrent loops through the cortex and basal ganglia to encode the serial order of two visual cues, A and B (Figure 3). The spiny neurons were simulated using the Gruber model [Gruber et al. 2003] with excitatory and postsynaptic inhibitory conductance inputs. Presynaptic inhibition was also modeled, by dynamically decreasing the excitatory synaptic weights. The GPi-T-PF loop (symbols in legend) was abstractly modeled based upon the Bieser and Houk model [1998] with a sigmoidal function to transform membrane potentials to firing rates. Response to a sequence of A followed by B is demonstrated (Figure 4). The network was then subjected to noise using no inhibition, presynaptic inhibition, and postsynaptic inhibition in caudate. A misclassification error in this example would be firing of the BA neuron in PF Cortex.

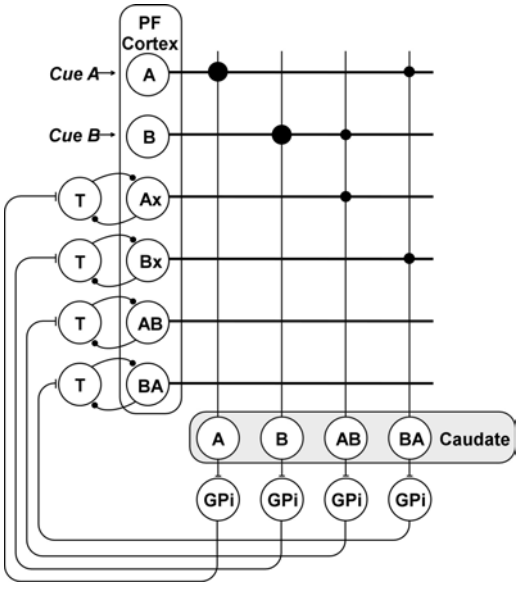


Figure 3: Serial order encoding network. Recurrent loops in the direct pathway through the prefrontal (PF) cortex, caudate (CD)

nucleus, globus pallidus pars internus (GPi), and thalamus (T) are used to encode two visual cues, A and B. Computational units AB and BA are labeled for the sequence they respond to best; Ax (Bx) is activated by A (B) independent of its serial order. Prefrontal cortex projections are excitatory, with synaptic weights represented by dot sizes. Caudate spiny units are interconnected by inhibitory collaterals to form a competitive network (shown symbolically by the shaded gray area). CD units are inhibitory to GPi, which in turn inhibit thalamic units. This disinhibition activates thalamic units. The loop is completed by reciprocal excitatory connections between thalamus and cortex.

Presynaptic inhibition yielded improved noise tolerance and decreased energy requirements compared to postsynaptic inhibition. When the network was subjected to noisy inputs, the misclassification rate without inhibition was 54.6% but fell to 24.1% for postsynaptic inhibition and 19.4% for presynaptic inhibition (4.8% decrease with presynaptic versus postsynaptic inhibition, $p<0.001$). Presynaptic inhibition also decreased the synaptic activity level in caudate from 118 to 98.0 (difference of 16.9%, $p<0.001$). This decreased excitatory synaptic activity may explain the reduced fMRI BOLD signal in caudate during the decoding contrast (Figure 2).

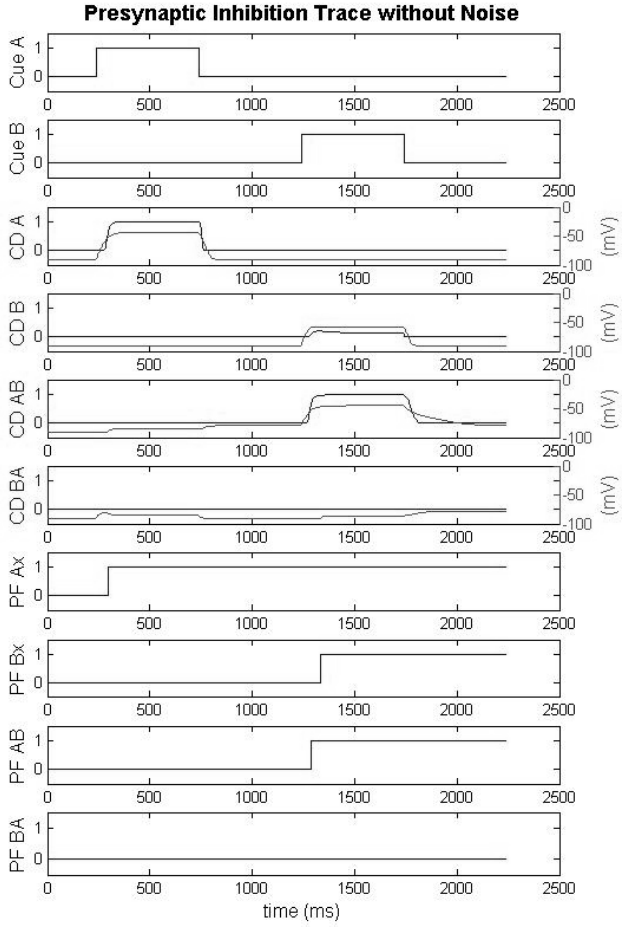


Figure 4: Response to a sequence of A followed by B using presynaptic inhibition in caudate. Firing rates in caudate (CD) and prefrontal cortex (PF) are on the left axis. CD membrane potentials are on the right axis. The effect of inhibition can best be seen as the membrane potential of the CD BA neuron is suppressed by lateral inhibition of the activated CD units. Inhibitory input from CD units causes tonically active GPi units to hyperpolarize and pause, producing rebound responses in thalamic units. The respective PF cortical units are then activated and sustained by positive feedback between thalamic and PF units.

3. Selection of Corrective Submovements

Tracking movements that require both speed and accuracy consist of a primary movement that is often off target, in which case it is followed by one or more corrective submovements in man [Novak et al. 2002] and in monkey [Fishbach et al. 2005]. The corrective submovements often overlap the primary movement, which suggests that the neural control system uses a forward model to predict the movement endpoint based on a copy of the neural command (efference copy) and a delayed sensory feedback. Whether the update of the neural command is continuous or intermittent is still under debate. Our findings from an analysis of the properties of submovements when perturbations of target location were introduced at movement onset strongly support the hypothesis that the neural controller predicts the need for a correction and selects an appropriate one intermittently, as illustrated in Figure 5.

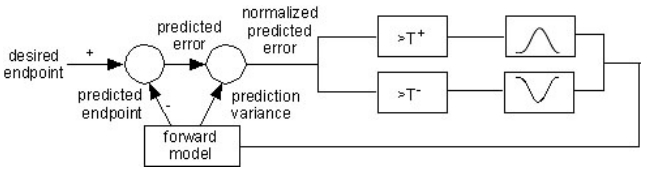


Figure 5: An operational model of how corrective submovements are generated. Vision provides the information about the desired endpoint, which can be updated as rapidly as 180 ms when a visual perturbation is introduced at movement onset. The brain computes the predicted endpoint based on efference copy and sensory input, and it computes the prediction variance based on past experience. The normalized predicted error (Z-score) must exceed a threshold value T in order to initiate a corrective submovement. The executed submovement follows an approximately bell-shaped velocity profile.

Single cell recordings in monkeys can be used to study how the basal ganglia participate in these processes. Since the output cells in GPi of the basal ganglia project to many different areas in the cerebral cortex, neurons need to be sampled from the region of GPi that projects to the primary motor cortex [Roy et al. 2003]. The sampled neurons should also be ones that are well related to the task. Figure 6 is an example that meets both of these criteria.

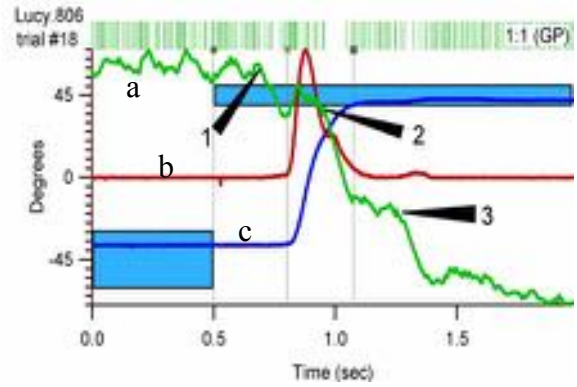


Figure 6: Activity during a single trial of a GPi neuron. In this task the monkey turns a rotating handle to move a cursor horizontally on a screen (blue trace (c) = position; red trace (b) = velocity) to acquire a target (boxes). The baseline-rate normalized cumulative sum histogram for the neuron ((a) green trace) shows three pauses in the high tonic discharge rate in this GPi neuron. The first pause (1) is small and occurs prior to the primary movement; the second and third are stronger pauses in association with tiny corrective submovements (2 & 3).

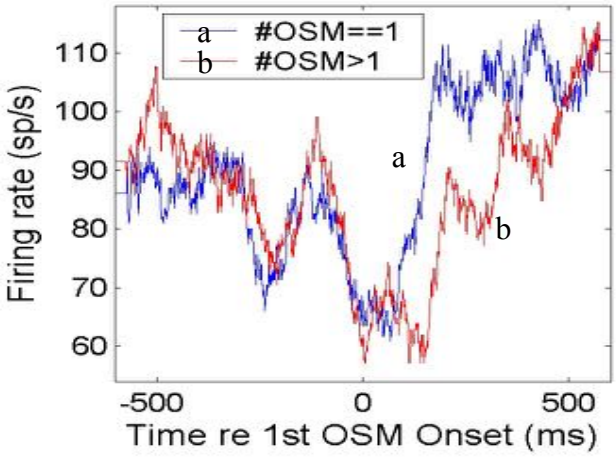


Figure 7 illustrates the reliability of single trial properties in a block of trials for the Figure 6 neuron. The average firing rate of the cell is shown for all trials containing a single corrective submovement ((a) blue trace) and for trials containing multiple corrective submovements ((b) red trace). Both traces are aligned to the onset of the first correction. Note that the pauses corresponding to the submovements are as strong or stronger than the pause for the primary movement, even though the corrections they appear to control are typically much smaller than is the primary movement. In the next paragraph, we discuss the likely explanation for these discrepant amplitude relationships.

The DPM model mentioned earlier (also see [Houk and Wise, 1995]), posits that practice in a task allows regularly rehearsed processing steps to be exported from the basal ganglia and/or cerebellum to the area of cerebral cortex to which the channel projects [Houk 2001, 2005]. Primary movements can be exported to the motor cortex since they are rehearsed in every trial. In contrast, the corrective submovements vary substantially from trial to trial, so nothing regular is rehearsed. This model of knowledge transfer from basal ganglia is supported by recent dual caudate-prefrontal cortex recordings of single cell activity [Pasupathy and Miller 2005].

4. Discussion

Our model of action selection is motivated by the extence of powerful computational features in the loops through the basal ganglia. The pattern classification operation shown in Figure 1 takes place in the striatal layer of a DPM. Computationally powerful pattern classification derives from several unique features of striatal medium spiny neurons [Houk 2005]. These features include: 1) a high convergence ratio [Kincaid et al. 1998] that presents nearly 20,000 different cortical inputs to any given spiny neuron, 2) a 3-factor learning rule that uses reward-predicting training signals from dopamine neurons to consolidate LTP learning [Houk et al. 1995], 3) an attentional neuromodulatory factor [Nicola et al. 2000] that induces bistability and nonlinear amplification in spiny neurons [Gruber et al. 2003], 4) competition among spiny neurons mediated by presynaptic and postsynaptic collateral inhibition [Figure 2; Plenz, 2003].

The anatomically demonstrated projections that loop back to the same area of cortex from which they derive [Kelly & Strick, 2004] allow cortical-basal ganglionic modules to perform serial order processing [Beiser and Houk 1998]. This feature allows them in principle to implement immediate serial order recall from working memories of a sequence. Long-term memories of serial order could be stored in cortico-cortical synapses or in the synapses between cortical neurons and striatal spiny neurons. The latter storage mechanism is thought to have a larger memory capacity for salient information [Houk & Wise, 1995]. Consistent with this hypothesis, the learning of new associations proceeds more rapidly in striatum than in cortex [Pasupathy & Miller, 2005]. The recall of previously learned sequences should also be efficient because cortical-basal ganglionic modules implement parallel searches through a vast repertoire of past experiences stored in the synapses of spiny neurons. Another important feature is that a network of DPM modules is in principle capable of recursion [Houk 2005], thus potentially resolving the "universal grammar" dilemma of language [Hauser et al. 2002].

Integrative control by basal ganglia and cerebellum. The present paper deals mainly with cortical-basal ganglionic loops whereas most DPMs also have loops through cerebellum. Regarding the latter, presently we know most about signal processing in the loops between cerebellum and primary motor cortex [Houk & Mugnaini 2002]. There are actually two loops in each cortical-cerebellar module. The one through the cerebellar nucleus is predominately excitatory and is responsible for the high firing rates of voluntary movement commands [Holdefer et al. 2005].

This is the amplification block in Figure 1 -- positive feedback is responsible for the amplification. The longer loop through cerebellar cortex uses the strong inhibitory output from Purkinje cells to restrain the positive feedback and, most importantly, to set the fixed points of this attractor network [Houk & Mugnaini 2002].

How do cortical-basal ganglionic and cortical-cerebellar modules work together? Figures 6 and 7 show an example of a GPi neuron in the basal ganglia helping to select a primary movement and subsequent submovements in a tracking task. These pauses will result in disinhibitions of the M1 neurons to which the GPi neuron, via thalamus, projects, thus facilitating one or more bursts of discharge. Each of these bursts would then need to be amplified and refined by the cerebellum. Amplification in intensity and time would serve to generate any given element of the M1 output vector in Figure 1, and spatial amplification would recruit the large population of M1 neurons (additional elements of that vector) which are required to produce a movement [Georgopoulos & Kristan, 2001]. The cerebellar cortex would then restrain and refine the entire M1 output vector, shaping it into a composite motor command calling for a primary movement and the subsequent corrective submovements that home in on the target.

The engineering operations in Figure 5 nicely superimpose on the neurophysiological operations abstracted in Figure 1. With the help of dopamine neuromodulation, pattern classification in the striatum should be able to generate the normalized predicted error in Figure 5, utilizing convergent cortical input reflecting both *phasic* sensory / efference copy events and *tonic* contextual cues. The 3-factor learning rule in striatum would, through prior experience, have stored these combined patterns in corticostriatal synaptic weights via reinforcement learning. The resultant output vector from the basal ganglia should then be able to embody appropriate motor cortical neurons for starting a movement in approximately the right direction, thus also initiating positive feedback and amplification in the loop through cerebellar nucleus. Finally, the Purkinje cells in the cerebellar cortex would shape population discharge into an output vector that commands a reasonable bell-shaped primary movement together with the subsequent corrective submovements that are needed to ensure an accurate overall movement.

Implications for schizophrenia. A simplified version of the Replicate task has been studied in patients suffering from schizophrenia [Fraser et al. 2004]. The patients exhibited two prominent deficits that were anticipated from existing models: (1) in line with predictions based on Monach's [2003] capacity model, serial order processing became saturated at 3 or 4 items in the list, as contrasted with the normal capacity of 7±2 [Miller 1956] and (2) in line with predictions based on the Beiser and Houk [1998] network model, targets presented later in the sequence were remembered most poorly. Both highly significant deficits were attributed to defective pattern classification in the caudate nucleus. This interpretation could be tested by imaging the Replicate task. One prediction to be tested is that the decrease in caudate blood flow in the decoding contrast (Figure 2) will be attenuated or even reversed in schizophrenia, assuming there is a deficit in GABAb mediated presynaptic inhibition.

In fact, there is a modified expression of the GABAb receptor in schizophrenia [Enna & Bowery, 2004]. This implicates the modified GABAbR1 gene on chromosome 6p21.3 [Martin et al. 2001] as a major contributor to schizophrenia. Since the inheritance of schizophrenia is multigenic [Freedman et al. 2001], the gene identified by Freedman, Leonard and collaborators is also strongly implicated, a gene that causes altered expression of a nicotinic receptor that is prevalent in many of the loops between the cerebral cortex and the cerebellar nuclei. Altered transmission in these loops is thought to contribute to the cognitive dysmetria of schizophrenia [Andreasen, 1999].

A central paradox of schizophrenia is that a condition which is genetic in origin survives in the population in spite of a substantial fecundity disadvantage. The magnitude of the latter is such that any genetic predisposition would be eliminated from the population within a few generations. Instead, since the incidence of schizophrenia remains steady at 1-2%, there must be an accompanying genetic advantage [Huxley et al. 1964]. In analyzing this issue, Kuttner et al. [1967] offered three potential advantageous functions that accompany the inheritance of schizophrenia: (1) a capacity for complex social relations, (2) intelligence and (3) language. Crow and colleagues have made a strong case for an evolutionary link between the origin of language and the etiology of schizophrenia [Berlim et al. 2003]. This hypothesis is consistent with the deficit in competitive pattern classification in schizo-phrenia mentioned earlier -- language contains abundant examples of serial order processing. One gene coding for GABAb receptors at presynaptic sites and another coding for nicotinic receptors, along with occasional malfunctioning variants associated with epigenetic expression, might explain the survival of genes responsible for schizophrenia.

5. Summary

We posit that both on-line error correction and serial order recall are examples of natural action selection. They appear to use analogous mechanisms for signal processing in their respective DPMs. Large-scale models comprised of interacting networks of DPMs may provide an ideal substrate for exploring the dynamics of the mind. Such simulations may also help us to understand schizophrenia.

Acknowledgement. This multimodal research was made possible by grant NS44837 from the National Institute of Neurological Disorders and Stroke.

References

Andreasen, N.C. (1999) A unitary model of schizophrenia: Bleuler's "fragmented phrene" as schizencephaly. *Arch. Gen. Psychiatry* 56: 781-787.

Arbib, M. A., A. Billard, M. Iacoboni and E. Oztop (2000). Synthetic brain imaging: grasping, mirror neurons and imitation. *Neural Netw* 13(8-9): 975-97.

Beiser, D. G. and J. C. Houk (1998). Model of cortical-basal ganglionic processing: encoding the serial order of sensory events. *Journal of Neurophysiology* 79: 3168-3188.

Berlim, M.T., B.S. Mattevi, P. Belmonte-de-Abreu, and T.J. Crow (2003) The etiology of schzophrenia and the origin of language: Overrview of a theory. *Comprehensive Psychiatry* 44: 7-14.

Botvinick, M. M. and D. C. Plaut (2005). Short-term memory for serial order: A recurrent neural network model. *Psychological Review*: in press.

Calabresi, P., N. B. Mercuri, M. DeMurtas and G. Bernardi (1991). Involvement of GABA systems in feedback regulation of glutamate- and GABA-mediated synaptic potentials in rat neostriatum. *Journal of Physiology* 440: 581-599.

Enna, S.J. and N.G. Bowery (2004) GABAb receptor alterations as indicators of physiological and pharmacological function. *Biochemical Pharmacology* 68: 1541-1548.

Fansler-Wald, D., A. Fishbach, D. Fraser, L. E. Miller and J. C. Houk (2004). Event sequence detection utilizing a minimal network of striatal spiny neurons. *Motor Systems Day*, Northwestern University.

Fishbach, A., S.A. Roy, C. Bastianen, L.E. Miller and J.C. Houk (2005). Kinematic properties of on-line error corrections in the monkey. *Exp. Br. Res.:* in press.

Fraser, D., S. Park, G. Clark, D. Yohanna and J. C. Houk (2004). Spatial serial order processing in schizophrenia. *Schizophrenia Research* 70(2-3): 203-213.

Freedman, R., S. Leonard, A. Oliney, C.A. Kaufmann, D. Malaspina, C.R. Cloninger, D. Svrakic, S.V. Faraone and M.T. Tsuang (2001). Evidence for the multigenic inheritance of schizophrenia. *Am. J. Medical Genetics (Neuropsychiatric Genetics)* 105: 794-800.

Georgopoulos, A. P. and W. B. Kristan (2001). "Motor System - Editorial overview." *Current Opinion in Neurobiology* 2001: 653-654.

Gruber, A. J., S. A. Solla, D. J. Surmeier and J. C. Houk (2003). "Modulation of striatal single units by expected reward: A spiny neuron model displaying dopamine-induced bistability." *J. Neurophysiology* 90: 1095-1114.

Gurney, K., T. J. Prescott and P. Redgrave (2001). "A computational model of action selection in the basal ganglia. I. A new functional anatomy." *Biol. Cybern.* 84: 401-410.

Hauser, M. D., N. Chomsky and W. T. Fitch (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* 298(5598): 1569-79.

Holdefer, R. N., J. C. Houk and L. E. Miller (2005). Movement-related discharge in the cerebellar nuclei persists after local injections of GABA-A antagonists. *J Neurophysiol* 93(1): 35-43.

Houk, J. (2001). Neurophysiology of frontal-subcortical loops. *Frontal-Subcortical Circuits in Psychiatry and Neurology.* D.G. Lichter and J. L. Cummings. New York, Guilford Publications: 92-113.

Houk, J. C. (2005). Agents of the Mind. *Biological Cybernetics* DOI 10.1007.

Houk, J. C., J. L. Adams and A. G. Barto (1995). A model of how the basal ganglia generates and uses neural signals that predict reinforcement. *Models of Information Processing in the Basal Ganglia.* J.C. Houk, J.L. Davis and D. G. Beiser. Cambridge, MA, MIT Press: 249-274.

Houk, J. and E. Mugnaini (2002). Cerebellum. *Fundamental Neuroscience.* L.R. Squire et al., Academic Press: 841-872.

Houk, J. C. and S. P. Wise (1995). Distributed modular architectures linking basal ganglia, cerebellum, and cerebral cortex: Their role in planning and controlling action. *Cerebral Cortex* 5: 95-110.

Huxley, J., E. Mayr, H. Osmond, A. Hoffer. (1964) Schizophrenia as a genetic morphism. *Nature* 204: 220-221.

Kelly, R. M. and P. L. Strick (2003). Cerebellar loops with motor cortex and prefrontal cortex of a nonhuman primate. *J. Neuroscience* 23: 8432-8444.

Kelly, R. M. and P. L. Strick (2004). Macro-architecture of basal ganglia loops with the cerebral cortex: use of rabies virus to reveal multisynaptic circuits. *Prog Brain Res* 143: 449-59.

Kincaid, A. E., T. Zheng and C. J. Wilson (1998). Connectivity and convergence of single corticostriatal axons. *J Neurosci* 18(12): 4722-31.

Kuttner, R.E., A.B. Lorincz, D.A. Swan. (1967) The schizophrenia gene and social evolution. *Psychol. Rep.* 20: 407-412.

Manoach, D.S. (2003) Prefrontal cortex dysfunction during working memory performance in schizophrenia: reconciling discrepant findings. Schizophrenia Research 60: 285 – 298.

Martin, S.C., Russek, S.J,, Farb, D.H. (2001) Human GABAbR genomic structure: evidence for splice variants in GABAbR1 but not GABAbR2. Gene 278: 63-79.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits to our capacity for processing information. *Journal of Experimental Psychology* 41: 329-335

Nicola, S. M., J. Surmeier and R. C. Malenka (2000). Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. *Annu Rev Neurosci* 23: 185-215.

Nisenbaum, E. S., T. W. Berger and A. A. Grace (1993). Depression of glutamatergic and GABAergic synaptic responses in striatal spiny neurons by stimulation of presynaptic GABA-B receptors. *Synapse* 14(3): 221-42.

Novak, K. E., L. E. Miller and J. C. Houk (2002). The use of overlapping submovements in the control of rapid hand movements. *Exp. Brain Res.* 144: 351-364.

Pasupathy, A. and E. K. Miller (2005). Different time courses of learning-related activity in the prefrontal cortex and striatum. *Nature* 433(7028): 873-6.

Plenz, D. (2003). When inhibition goes incognito: feedback interaction between spiny projection neurons in striatal function. *TINS* 26: 14427–14432.

Raymond, J. L., S. G. Lisberger and M. D. Mauk (1996). The cerebellum: A neuronal learning machine? *Science* 272: 1126-1131.

Redgrave, P., T. J. Prescott and K. Gurney (1999). The basal ganglia: A vertebrate solution to the selection problem? *Neuroscience* 89(4): 1009-1023.

Roy, S. A., C. Bastianen, E. Nenonene, A. Fishbach, L. E. Miller and J. C. Houk (2003). Neural correlates of corrective submovement formation in the basal ganglia and motor cortex. *Society for the Neural Control of Movement Abstracts*.

Rubchinsky, L. L., N. Kopel and K. A. Sigvardt (2003). Modeling facilitation and inhibition of competing motor programs in basal ganglia subthalamic nucleus–pallidal circuits. *PNAS* 100: 14427–14432.

Tepper, J. M., T. Koos and C. J. Wilson (2004). GABAergic microcircuits in the neostriatum. *Trends Neurosci* 27(11): 662-9.

Cognition, Action Selection, and Inner Rehearsal

Murray Shanahan

Department of Electrical & Electronic Engineering,
Imperial College London,
Exhibition Rd., London SW7 2BT.

Abstract

This paper presents a large-scale model of the architecture of the mammalian brain, the core circuit of which carries out inner rehearsal of interaction with the environment to realise a form of cognitively mediated action selection. As it alternates between broadcast to and competition between its component neural assemblies, the core circuit exhibits an episodic dynamics suggestive of cortical processing in discrete frames. The implemented architecture is used to control a simulated robot, and a classic experimental paradigm in which rats performed apparently goal-directed action selection is emulated.

1 Introduction

In the 1940s, Tolman and Gleitman used a classic experimental setup to demonstrate apparently goal-directed behaviour in rats (Tolman & Gletiman, 1949). The rats were allowed to explore a T-maze containing a dark room on the left and a light room on the right (Fig. 1, left). Both rooms contained food. The rats were then placed in a separate enclosure resembling the dark room, and subjected to electric shocks through the feet. When reintroduced to the base of the T-maze, the rats always navigated directly to the light room, even though the actions of turning left and right had been equally reinforced.

The rat's ability to "think ahead" in this situation is hard to explain using reinforcement alone, and seems to require the inference of an indirect cause-and-effect relationship. However, Hesslow (2002) argues that the only extension to the paradigm of classical conditioning required to explain this sort of behaviour is a mechanism for inner rehearsal. Indeed, both Cotterill (1998) and Hesslow (2002) propose internally simulated interaction with the environment as the very basis of animal and human cognition.

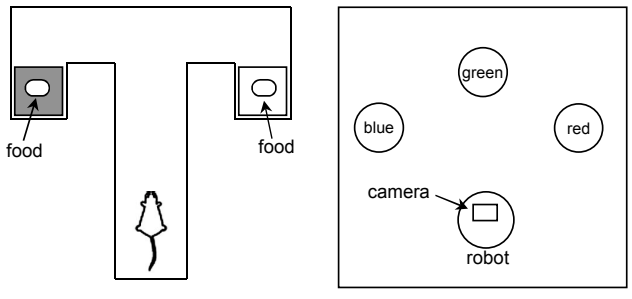


Fig 1: Rat and robot experiments

In pursuit of this suggestion, the present paper describes a large-scale, high-level neural model that realises goal-directed action selection for a simulated robot in an analogous experimental setup (Fig. 1, right). The model implements an architecture whose core circuit carries out inner rehearsal to anticipate the effects of currently executable actions, which are held on veto while these anticipated effects are evaluated by an affective system. This can bring about an increase or decrease in an action's salience, which in turn can result in the strengthening or weakening of its veto. When an action's salience exceeds a given threshold, its veto is released and the action is carried out.

The design of the core circuit facilitates the integration of the activities of multiple, parallel neural assemblies using a combination of competition and broadcast, and thereby realises a global workspace architecture (Baars, 1988; 2002). The dynamics of the core circuit exhibits a pattern of alternation between stability and rapid change, and is reminiscent of certain recent EEG findings suggestive of the idea that the cortex processes information in discrete frames (Freeman, 2003; 2004).

2 The Architecture of the Model

Fig. 2 shows a top-level schematic of the model's architecture. It can be thought of in terms of two interacting subsystems. The first-order system is purely reactive, and determines an immediate motor response to the present situation without the intervention of cognition. But these unmediated motor responses are subject to a veto imposed by BG (the basal ganglia analogue). Through BG, which carries out salience-based action selection, the higher-order loop modulates the behaviour of the first-order system. It does this by adjusting the salience of currently executable actions. Sometimes this adjustment will result in a new action becoming the most salient, and sometimes it will boost an action's salience above the threshold required to release its veto, bringing about that action's execution.

The higher-order system determines these salience adjustments by carrying out off-line rehearsals of trajectories through (abstractions of) the robot's sensorimotor space. In this way – through the exercise of its "imagination" – the robot is able to anticipate and plan for potential rewards and threats without exhibiting overt behaviour. The first- and higher-order systems have the same basic components and structure. Both are sensorimotor loops. The key difference is that the first-order loop is closed through interaction with the world itself while the higher-order loop is closed internally. This internal closure is facilitated by AC, which simulates — or generates an abstraction of — the sensory

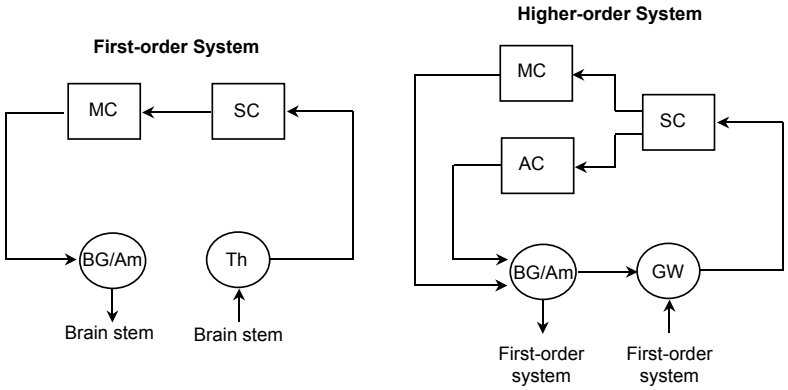


Fig. 2: A top-level schematic of the architecture. MC = motor cortex, SC = sensory cortex, AC = association cortex, BG = basal ganglia, Am = amygdala, Th = thalamus.

stimulus expected to follow from a given motor output, and fulfils a similar role to that of a *forward model* in the work of various authors (Demiris & Hayes, 2002; Hoffman & Möller, 2004; Grush, 2004; Ziemke, *et al*., 2005). The cortical components of the higher-order system (SC, AC, and MC) correspond neurologically to regions of association cortex, including the prefrontal cortex which is implicated in planning and working memory (Fuster, 1997).

2.1 Affect and Action Selection

Analogues of various sub-cortical and limbic structures appear in both the first- and higher-order systems, namely the basal ganglia, the amygdala, and the thalamus. In both systems, the basal ganglia are implicated in action selection. Although, for ease of presentation, the schematic in Fig. 2 suggests that the final stage of motor output before the brain stem is the basal ganglia, the truth is more complicated in both the mammalian brain and the robot architecture under discussion.

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

The robot architecture includes analogues of the basal ganglia and amygdala that function in a similar way. These operate in both the first- and higher-order systems. In the first-order system, the amygdala analogue associates patterns of cortical activation with either reward or punishment, and thereby modulates the salience attached to each currently executable action (Balkenius & Morén, 2001). The basal ganglia analogue adjudicates the competition between each executable action and, using a winner-takes-all strategy, selects the most salient for possible execution (Prescott, *et al*., 1999). While the salience of the selected action falls below a given threshold it is held on veto, but as soon as its salience exceeds that threshold it is executed.

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

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

2.2 Global Workspace Theory

An important feature of the architecture, though not one that is explored fully in the present paper, is that it conforms to global workspace theory (Baars, 1988), which advances a model of information flow in which multiple, parallel, specialist processes compete and co-operate for access to a global workspace. Gaining access to the global workspace allows a winning coalition of processes to broadcast information back out to the entire set of specialists. Although the global workspace exhibits a serial procession of broadcast states, each successive state itself is the integrated product of parallel processing.

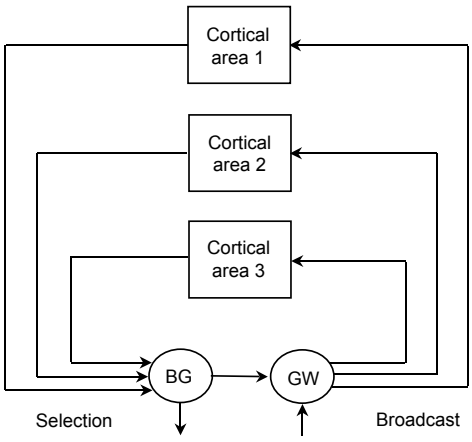


Fig 3: The fan-and-funnel model

According to global workspace theory, the mammalian brain instantiates this model of information flow, which permits a distinction to be drawn between conscious and unconscious information processing. Information that is broadcast via the global workspace is consciously processed while information processing that is confined to the specialists is unconscious. A considerable body of empirical evidence in favour of this distinction has accumulated in recent years (Baars, 2002). Although the topic of consciousness is orthogonal to the present paper, the combination of broadcast and competition that is the hallmark of the global workspace architecture is central to the action selection mechanism under investigation. During the process of internally exploring a space of possible sensorimotor trajectories, broadcast enables multiple branch points to be considered – in effect engaging many forward models simultaneously – while competition determines which of the candidate branches is actually explored next.

Moreover, the particular blend of serial and parallel computation favoured by global workspace theory suggests a way to address the frame problem – in the philosopher's sense of that term (Fodor, 2000) – which in turn suggests that conscious information processing may be cognitively efficacious in a way that unconscious information processing is not (Shanahan & Baars, 2005). In particular, in the context of so-called informationally unencapsulated cognitive processes, it allows relevant information to be sifted from the irrelevant without incurring an impossible computational burden. More generally, broadcast interleaved with competition facilitates the integration of the activities of large numbers of specialist processes working separately. So the global workspace model can be thought of as one way to manage the massively parallel computational resources that surely underpin human and animal cognitive prowess.

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

A number of candidate structures exist in the brain that might fulfill the role of GW. For example, the first-order / higher-order distinction is preserved in the thalamus, which contains not only first-order relays that direct signals from the brain stem up to cortex (located, for example, in the lateral geniculate nucleus), but also higher-order relays that route cortical traffic back up to cortex (located, for example, in the pulvinar) (Sherman & Guillery, 2001). So the thalamus is one plausible candidate for a broadcast mechanism in the mammalian brain. But the same function could be reliased by long-range corticocortical fibres, as proposed by Dehaene, *et al*. (2003), or indeed by some combination of thalamocortical and corticocortical communication.

Thankfully, there is no need to take a stand on this issue to supply an explanatory framework at an architectural level. What matters more in the present context is that the fan-and-funnel model of broadcast / distribution and competition / selection can be straightforwardly combined with the top-level schematic of Fig. 2, as is apparent from the diagrams. Indeed, the role of the BG component of the

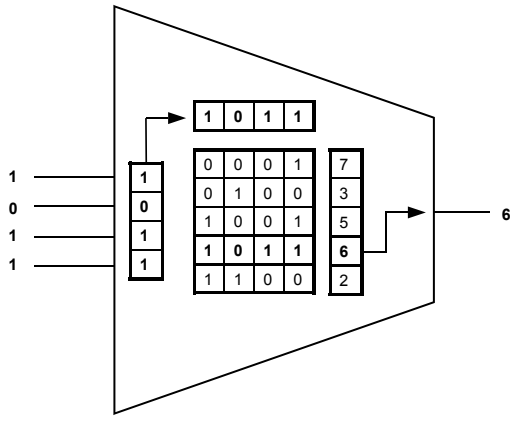


Fig 4: The G-RAM weightless neuron

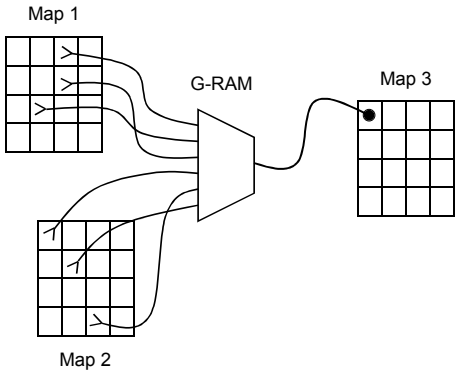


Fig 5: G-RAM maps and connections

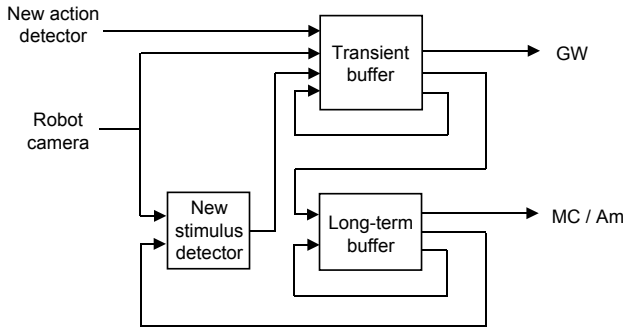


Fig. 6: Visual system circuitry (VC / IT). VC = visual cortex, IT = inferotemporal cortex.

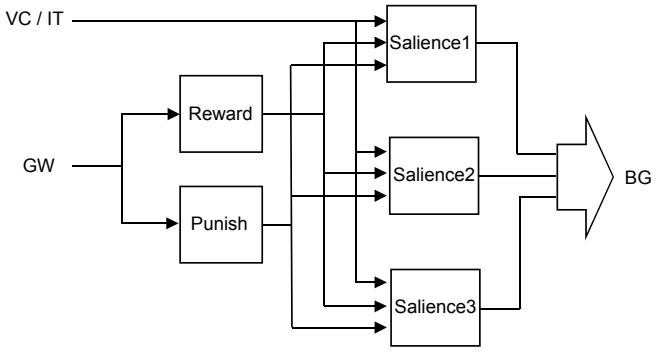


Fig. 7: Affect circuitry (Am)

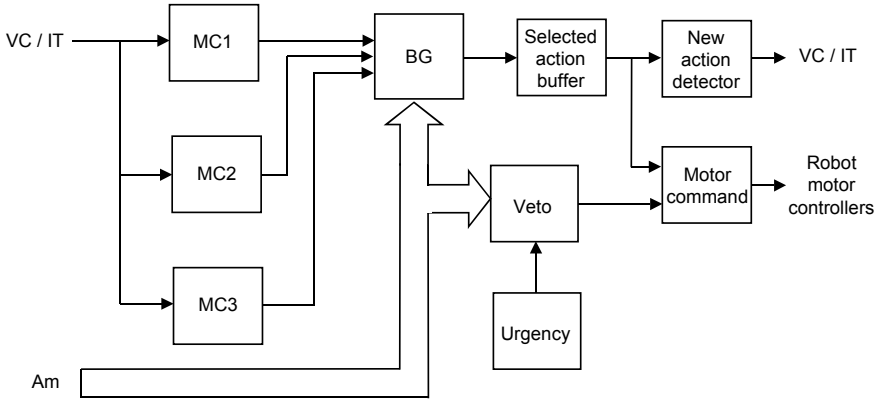


Fig. 8: Action selection circuitry (BG / MC)

higher-order loop introduced in Fig. 2 is precisely to effect the sort of selection between the outputs of multiple competing cortical areas shown in Fig. 3.

3 An Implementation

The brain-inspired architecture of the previous section has been implemented using NRM, a tool for building large-scale neural network models using G-RAMs (generalising random access memories) (Figs. 4 and 5). These are weightless neurons employing single-shot training whose update function can be rapidly computed (Aleksander, 1990), and which can be easily organised into attractor networks with similar properties to Hopfield nets (Lockwood & Aleksander, 2003).

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

The core of the implementation, which comprises almost 40,000 neurons and over 3,000,000 connections, is a set of cascaded attractor networks corresponding to each of the components identified in the architectural blueprint of the previous section. The NRM model is interfaced to Webots, a commercial robot simulation environment. The simulated robot is a Khepera with a 64 × 64 pixel camera, and the simulated world contains cylindrical objects of various colours. The Khepera is programmed with a small suite of low-level actions including "rotate until an object is in the centre of the visual field" and "approach an object in the centre of the visual field". These two actions alone are sufficient to permit simple exploration and navigation in the robot's simple environment.

The overall system can be divided into four separate modules – the visual system (Fig. 6), the affective system (Fig. 7), the action selection system (Fig. 8), and the broadcast / inner rehearsal system (Fig. 9). Each box in these figures denotes a layer of neurons and each path denotes a bundle of connections. If a path connects a layer A to an $n \times n$ layer B then it comprises n^2 separate pathways – one for

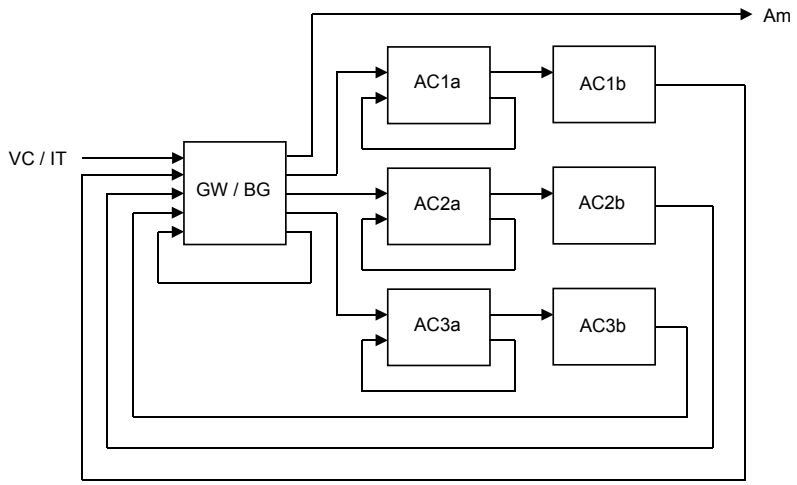


Fig. 9: Circuitry for broadcast and inner rehearsal (GW / BG / AC). GW = global workspace.

each of the neurons in *B* – each of which itself consist of *m* input connections originating in a randomly assigned subset of the neurons in *A* (Fig. 5). For the majority of visual maps *m* is set to 32.

The two buffers in the visual system comprise 64 × 64 topographically organised neurons (Fig. 6). These are both attractor networks, a property indicated by the presence of a local feedback path. The transient buffer is activated by the presence of a new visual stimulus. The hallmark of a new stimulus is that it can jog the long-term visual buffer out of one attractor and into another. The GW component of the inner rehearsal system is loaded from the transient visual buffer, whose contents rapidly fade allowing the dynamics of inner rehearsal to be temporarily dominated by intrinsic activity rather than sensory input.

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

So far the mechanism described is little different from a standard behaviour-based robot control architecture. What sets it apart from a purely reactive system is its capacity for inner rehearsal. This is realised by the core circuit depicted in Fig. 9, which is similar in both structure and function to the recurrent neural network of Tani (1996). When a new visual stimulus arrives, it overwrites the present contents of GW, and is thereby broadcast to the three cortical association areas AC1a to AC3a. The contents of these areas stimulates the association areas AC1b to AC3b to take on patterns of activation corresponding to the expected outcomes of the actions recommended by their motor-cortical counterparts. These patterns are fed back to GW / BG, leading to further associations corresponding to the outcomes of later hypothetical actions. By following chains of associations in this way, the system can explore the potential consequences of its actions prior to their performance, enabling it to anticipate and plan ahead.

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

To ensure that the system never gets stuck in a "thinking rut", endlessly pondering the possible consequences of its actions instead of actually doing something, a small population of neurons acts as an indicator of the urgency with which the robot should act (Fig. 8). At the onset of a new stimulus, this neural population becomes quiescent, reflecting a lack of urgency, holding the currently selected action on veto and giving the inner rehearsal system time to work. But its level of activity grows with time, reflecting an increasing sense of urgency, and the need to act soon. The veto on the execution of the currently most favoured action is thereby gradually weakened, and eventually this action

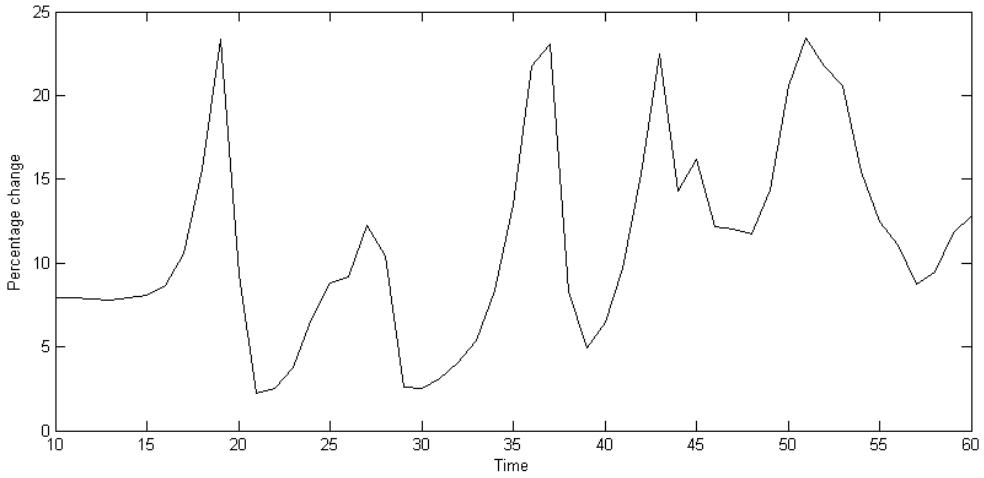


Fig. 10: Cycles of stability and instability

will be executed regardless of ongoing rehearsal. In this way, a balance is struck between reactivity and cognitively mediated, deliberative behaviour.

4 Results and Discussion

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

In each of the following experiments, the system runs a predefined training script prior to exhibiting the behaviour reported. Running this script sets up associations between patterns of visual input (VC / IT) and, for a subset of the three motor-neuronal assemblies (MC1 to MC3), corresponding recommended actions (Fig. 8) and their saliences (Fig. 7). This is analogous to reinforcement learning, acquiring a number of preferred immediate responses to an ongoing situation. In addition, the training script sets up associations between the current contents of GW and the punishment / reward neurons of Fig. 7. These permit the inner rehearsal mechanism, via the amygdala (Am), to exercise its influence on action selection. Producing similar results with a less supervised form of learning is an obvious theme for future research.

Fig. 10 illustrates an interesting property of the circuit of Fig. 9. The graph plots the percentage of neurons in the four maps GW and AC1a to AC3a that changed state from one time step to the next (where a time step corresponds to one complete cycle of updates to all the neurons in the system) during a typical run in which no external sensory input was presented to the robot. (A similar pattern is typically produced soon after the initial presentation of an external stimulus.) In order to study long chains of associations, a set of images of abstract coloured shapes (lozenges, stars, and so on) was used as a training set, rather than images obtained from the Webots simulator. But the same effect is apparent with images obtained directly from the simulated robot's camera. Specifically, the graph shows that the system of inner rehearsal exhibits a procession of stable states punctuated by episodes of instability, a pattern which is reminiscent of the recently reported phenomenon of aperiodic alternation between pan-cortical coherent and decoherent EEG activity (Freeman & Rogers, 2003; Freeman, 2004). According to Freeman, these results suggest that the cortex processes information in a series of movie-like frames corresponding to "recurring episodes of exchange and sharing of perceptual information among multiple sensory cortices" (Freeman, 2004, p. 2077).

In a similar vein, the periods of stability depicted in the graph occur when the contents of GW is being successfully broadcast to the three cortical regions, while the spikes of instability indicate that GW is being nudged out of its previous attractor and is starting to fall into a new one. The new attractor will be the outcome of a competition between AC1b to AC3b. The resulting new contents of GW is then broadcast to AC1a to AC3a, causing new activation patterns to form in AC1b to AC3b, which in turn give rise to a renewed competition for access to GW. This tendency to chain a series of associations together is what gives the system its ability to look several actions ahead.

Tables 1 and 2 summarise episodes within two typical runs of the system, corresponding respectively to the without-aversion and with-aversion conditions in the classic experiment of Tolman & Gleitman (1949) described in the introduction (Fig. 1, left). Each episode starts with the initial presentation of a new stimulus, and ends with the robot's first action. Under both conditions, the robot's environment contained just three cylinders – one green, one red, and one blue (Fig. 1, right). Area MC1 of the motor-cortical system was trained to recommend "rotate right" (RR) when presented with a green cylinder, while area MC2 was trained to recommend "rotate left" (RL).

Table 1: Without aversion to red cylinders

Time	Events
0	Green cylinder comes into view.
2	Green cylinder image in both visual buffers. MC1 recommends RR, MC2 recommends RL. RR has higher salience and is currently selected action. Veto is on.
3	Green cylinder image in GW and broadcast to AC1a to AC3a. AC1b has association with red cylinder, AC2b has association with blue cylinder.
6	Associated red cylinder image in GW.
8	Affective system quiescent, but urgency increasing.
19	Urgency very high. Veto released.
20	RR passed on to motor command area. Robot rotates right until red cylinder in view.

Table 2: With aversion to red cylinders

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

The action selection networks were trained in such a way that MC1's recommendation (rotate right) had the higher initial salience, and in a purely reactive system this action would have been immediately executed under both the without- and with-aversion conditions. But thanks to the imposition of a veto, the inner rehearsal system had a chance to anticipate the outcome of the recommended action, giving rise to contrasting behaviours in the two experimental conditions, as in Tolman and Gleitman's rat experiments. The inner rehearsal system was trained, using a predefined script matching the experimental setup, to associate 1) the RR action and the image of the green cylinder with the subsequent presentation of the red cylinder, and 2) the RL cylinder and the image of the green cylinder with the subsequent presentation of the blue cylinder.

To emulate the without-aversion condition, the affective system was trained so that neither its "reward" nor its "punishment" neurons fired when GW contained the image of a red cylinder. Under this condition, the robot's behaviour is the result of pure reinforcement. As Table 1 shows, this brought about the execution of RR – the system's immediately preferred, reactive response – as soon as the combination of urgency and salience exceeded the threshold required to release the veto on that action.

By contrast, to emulate the with-aversion condition, the "punish" neurons were trained to fire when GW contained the image of the red cylinder. As Table 2 shows, this lead the system to reduce the salience of its initially preferred action (RR) following a period of inner rehearsal that revealed its unpleasant expected consequences. The inner rehearsal system then explored the consequences of the alternative RL action. When these turned out to be more palatable, the salience of the RL action increased until its veto was eventually released, the RL command was forwarded to the motor output area, and the robot finally rotated to face the blue cylinder.

As all of this took place, urgency was increasing, but not fast enough to outpace the process of rehearsal and prevent it from influencing the selected action. The upper row of Table 3 summarises the results of eight further trials under the with-aversion condition, using the same training script but with a different randomly generated network configuration for each trial. The RL action is selected on each occasion, with some variation in timing.

Table 3: Sample runs with aversion

	Time to first action / action taken							
	1	2	3	4	5	6	7	8
μ=8	17 RL	16 RL	15 RL	16 RL	15 RL	15 RL	20 RL	15 RL
μ=24	6 RR	9 RL	3 RR	15 RL	2 RR	14 RL	17 RL	3 RR

The behaviour the system exhibits under these two experimental conditions demonstrates that the architecture is capable of an elementary from of cognitively mediated action selection similar to that first reported by Tolman and Gleitman (1949). Moreover, the architecture is broadly consistent with contemporary high-level neuroanatomy, and it conforms to the theoretical proposals of both Baars (1998) and Hesslow (2003). In addition, the episodic dynamics of its core circuit is supportive of Freeman's interpretation of recent EEG findings in terms of discrete frames of cortical processing (Freeman & Rogers, 2003; Freeman, 2004). Neither the architecture nor the current implementation is confined to the simple experimental setup described in this paper, and their use in richer environments is the subject of ongoing work.

For example, by varying the system's baseline level of urgency (μ), it is possible to adjust the trade-off between

deliberation and reactivity – a high baseline level of urgency results in a tendency to act quickly but "unthinkingly" (Table 3, lower row), while a low baseline level of urgency results in slower but sometimes more effective action selection (Table 3, upper row). Preliminary experimentation also suggests that it may be possible to reproduce the behavioural phenomenon of "microchoices" reported by Brown (1992), wherein rats make tentative small explorations of arms of a star-maze before eventually making an apparently goal-directed choice. Using the mechanisms described here, a similar effect can be had by selecting a baseline level of urgency that allows for some anticipation of the consequences of actions, but only enough to look a very few actions ahead. The long-term hope is that, through experiments such as this, the conceptual framework and architecture of the present paper will help to further our understanding of the basis of cognition in both animals and machines.

Acknowledgments

Thanks to Igor Aleksander, Bernie Baars, Rodney Cotterill, Barry Dunmall, Gerry Hesslow, and to the workshop's anonymous reviewers.

References

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

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

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

Balkenius, C. & Morén, J. (2001). Emotional Learning: A Computational Model of the Amygdala. *Cybernetics and Systems 32* (6), 611–636.

Baxter, M.G. & Murray, E.A. (2002). The Amygdala and Reward. *Nature Reviews Neuroscience 3*, 563–573.

Brown, M.F. (1992). Does a Cognitive Map Guide Choices in the Radial Arm Maze? *Journal of Experimental Psychology: Animal Behavior Processes 18* (1), 56–66.

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

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

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

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

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

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

Freeman, W.J. (2004). Origin, Structure, and Role of Background EEG Activity. Part 1. Analytic Amplitude. *Clinical Neurophysiology 115* (9), 2077–2088.

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

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

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

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

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

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

Prescott, T.J., Redgrave, P. & Gurney, K. (1999). Layered Control Architectures in Robots and Vertebrates. *Adapative Behavior 7* (1), 99–127.

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

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

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

Tani, J. (1996). Model-Based Learning for Mobile Robot Navigation from the Dynamical Systems Perspective. *IEEE Transactions on Systems, Man, and Cybernetics B 26* (3), 421–436.

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.

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

Goal and motor action selection using a hippocampal and prefrontal model

Nicolas Cuperlier, Philippe Gaussier, Philippe Laroque, Mathias Quoy
ETIS-UMR 8051
Universit de Cergy-Pontoise - ENSEA
6, Avenue du Ponceau
95014 Cergy-Pontoise France
cuperlier@ensea.fr

Abstract

We have developped a mobile robot controller based on hippocampus and prefrontal models. The model adresses two action selection problems encoutered in navigation tasks: selection of the goal (in a the case of multiple and contradictarory goals) and choice of the motor actions according to the local situation. It relies on a cognitive map linking "transition cells" coding for the transition between two places successively recognized. We propose these transitions are learned and predicted by a simple neural mechanism corresponding to the hippocampus. Each transition cell can then be associated with the integrated direction used during the displacement. When several contradictory transitions are possible, a small bias from the planning system (prefrontal cortex) is sufficient to select/filter the appropriate transitions. Final selection of the motor action results from the merging of these global decisions with local constraints such as obstacle avoidance, robot inertia... We show a dynamical neural field is a simple and efficient solution to solve these possible contradictions and allow a stable and correct behavior. Simulations and robotics experiments are used to illustrate theses mechanisms.

1 Introduction

Path planning requires from the agent or the robot to select the appropriate action to perform. This task might be complex when several actions are possible, and so different approaches have been proposed to choose what to do next.

Some works use ruled-based algorithms, classical functional approach, that can exhibit the desired behaviors, we will not discuss them in this paper, but one can refer to [Donnart and Meyer, 1996; Tyrrell, 1993]. Instead, other works try to look at what the nature does by taking inpiration from neurobiology to design control architecture. There are at least two reasons for this:

- first, getting robust, adaptative, opportunistic and ready-made solutions for control architecture.
- second, if robotic results can be compared to experimental results involving several parts of the brains which are generally difficult to study due to its complexity, it can help neurobiologist to understand how a neurobiological model behaves.

Experiments carried out on rats have led to the definition of cognitive maps used for path planning [Tolman, 1948]. Most of cognitive maps models are based on graphs showing how to go from one place to an other [Arbib and Lieblich, 1977; Samsonovich and McNaughton, 1997; Bachelder and Waxman, 1994; Trullier *et al.*, 1997; Schölkopf and Mallot, 1995; Bugmann *et al.*, 1995]. They mainly differ in the way they use the map in order to find the shortest path, in the way they react to dynamical environment changes, and in the way they achieve contradictory goal satisfactions. We will focus here on models inspired by the possible use of particular neurons of the rat's hippocampus, called *place cells* [O'Keefe and Nadel, 1978]. These neurons fire when a rat is at a particular location in its environment. We will first present the simulated environment and the animat possible behaviors (section 2), then we will describe our model for goal selection, based on transition cells (section 3). Finally, we will use a neural field selecting the final movement to perform (section 4). Simulations are carried out on an *animat*, and real world experiments on a Labo3 *robot*.

2 Animat behaviors

The models studied in this paper have all been experimented using an approach inspired from the concepts of situated agents and animats [Meyer and Wilson, 1991]. We suppose our animats live in an unknown environment with several sources (like "food", "water" and "nest") and some obstacles. We use three contradictory motivations (eating, drinking, and resting) each one associated with a satisfaction level that decreases over time and increases when the animat is on the proper source. We do not provide any ad hoc description of the environment. Indeed, the animat gets only two types of information: the presence of landmarks from its visual input and the azimuth of these landmarks relatively to the north given by a compass or a vestibular system [Arleo and Gerstner, 2000]. Animats have four possible behaviors for deciding which action to realize. They are given here in decreased order of priority:

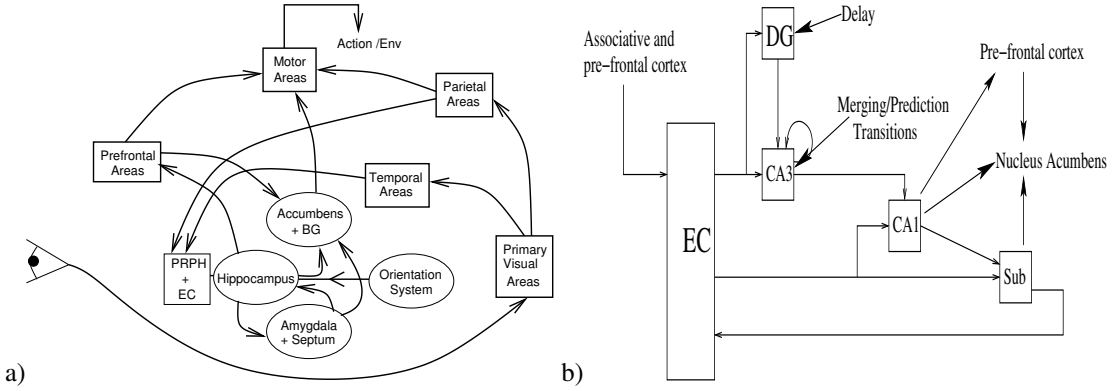


Figure 1: a)Schematic figure of the brain structures we are intersted in. PRPH represent the perirhinal and the parahippocampus. b)Details of the sub-structures of the hippocampus.

1. Random exploration to discover the environment and the needed sources.
2. Planning to reach the sources in order to satisfy the animat's motivations.
3. Obstacles avoidance allows the animat to follow obstacles until the desired movement becomes possible.
4. When the source is very near, the animat directly sees where the source is located and uses this information to reach it.

3 Goal selection model

Our model focuses on a loop formed by the hippocampus (HS), the prefrontal cortex (PF) and the basal ganglia (BG). These two last structures are modelized from a functionnal approach rather than in a detailed manner (see forward)).

The hippocampus is a brain sub-cortical structures which takes input from the whole associative areas (see fig.1) via the entorhinal cortex (EC) and projects efferences into associative, prefrontal and premotor cortical areas. A schematic overwiew of the structure we are interested in is given in figure 1. It has been shown that the hippocampus was involved when performing a navigation task since place cells have been found in the rat's hippocampus (particularly CA3, CA1 and DG regions) and in the entorhinal cortex (EC) [O'Keefe and Nadel, 1978].

There is no need for a Cartesian map since a particular place is defined by a given set of (landmark, azimuth) pairs. The recognition of the present location is based on the landmark configuration. A place cell P_c responds according to the position of the animat in its environment. In our model its activity is calculated as the distance between a landmark configuration learned and the present one. So, the higher this response, the closer the animat is to P_c. After competition between all place cells, the winning cell represents the location where the animat thinks it is [Gaussier *et al.*, 2000]. As a place cell still keeps a certain amount of activity even if the animat isn't near the coded place, the neuron place filed can be quite large. Consequently, we use a rule that controls the recruitment of a new neuron. Hence, a new neuron will code for a place, if all previously learned neurons have an activity lower than the Recognition Threshold (RT). A place cell may be linked with the movement needed to reach a goal. This sensory-motor association may be generalized to the whole environment [Gaussier *et al.*, 2000]. However, this simple reactive mechanism is not enough in environments composed of several rooms, or when there are contradictory motivations. A cognitive map will solve these drawbacks. Transition cells are

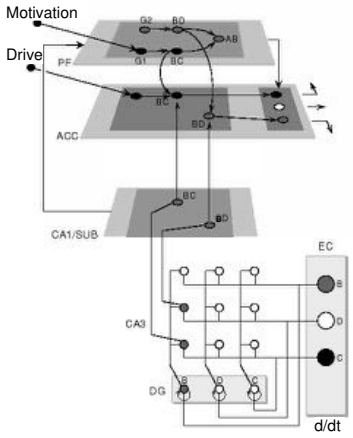


Figure 2: Planning from transition cells. Place cells recognitions feeds the transition prediction mechanism that provides all possible transitions beginning with the corresponding place cells. In this example, transitions BC and BD are predicted since two different action has been learned from B location. The choice of the correct transition to use is performed (in ACC) by the bias given by the activity of the goal level transitions (PF). As a higher activity means a shorter path to the goal, the BC transition is selected, since the goal is located at location C.

inspired by a neurobiological model of timing and temporal sequence learning in the hippocampus [Banquet *et al.*, 1997; Gaussier *et al.*, 2002; Banquet *et al.*, 2005].

Figure 2 shows the hippocampal model and the cognitive map in the prefrontal cortex. Place cells are created in EC by learning the landmark-azimuth configuration. Transition cells

are formed in CA3 by the merging of the current location in EC and the previous one in DG.

The transition cell is also coded in CA1. A path integration mechanism computes the mean direction used for going from one place to the other [Samsonovich and McNaughton, 1997]. This direction is linked with the transition at the output of the nucleus accumbens (ACC).

The cognitive map is located in our model in the prefrontal cortex (PF). The prefrontal cortex, in our model, is a simple neural network encoding a graph. It is built by linking transition cells successively reached during exploration and seems to be coherent with neurobiological data [V. Hok and Poucet, to appear in 2005]. Learning the cognitive map is performed continuously (latent learning). There is no separation between the learning and planning phase. This graph is a topological representation of the explored environment. The transition cells are the nodes of this graph and synapses link the transitions successively reached. Each source place is associated with a motivation neuron. This allows to define a road to be followed for reaching the goal: the activity diffuses along the links on the map and activates transition cells according to their distance (in number of links) to the goal. Diffusion is achieved in a way resulting to similar results than the Bellman-Ford one [Bellman, 1958]. This activity is sent to ACC layer where it is added to the proposed transitions. This bias allows the selection of the most activated transitions via a competition mechanism. Finally the corresponding movement is triggered (see fig. 3).

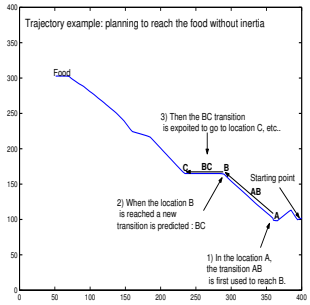


Figure 3: The animat first follows the direction coded by B and when it comes in B, the transition BC is predicted, and the corresponding direction is used.

A relevant question is about the growth of the number of transition cells created while exploring the environment. In order to answer this question we first have to underline that this number is intimately linked with the number of place cells, and above all, the number of place cells created for a fixed RT value, depends on the complexity of the environment. The degree of complexity of an environment relies mainly on two factors: the number and the location of its landmarks and the number of obstacles found inside. Hence, we have studied the ratio between created transition cells over created place cells for three environments of increasing complexity according to their obstacle configuration. For these tests, we have chosen to set the number of landmarks at a high value. For each experiment, we have launched a series of animats until 10 survive, and we let them live for 50000 cycles. This number has been chosen high enough to be sure that the animat has learned a complete cognitive map of the environment. The results shown here are the average on these 10 animat results. We have done these tests for a single, a two and a four room environment. The ratio remains stable around the mean value 5.45 for all environments once the cognitive map of the environment is complete (see table 1). Indeed, only a few transitions can be created, since a transition is a link between "adjacent" place cells. Furthermore the number of a place cell neighbours is necessary limited. So there is no combinatorial explosion on the number of created transitions.

Env / RT	0.97
nbp	133.8(2.85)
nbt	735.8(19.80)
ratio	5.49(0.06)
nbp	606.2(6.89)
nbt	3389.2(56.38)
ratio	5.59(0.08)
nbp	643.7(9,88)
nbt	3281,2(48,80)
ratio	5.09(0,04)

Table 1: Ratio of the number of place cells (nbp) created over the number of transitions created (nbt) according to the number of room in the environment: with one room (top line), with two rooms (middle line) and four rooms (bottom line). Standard deviation is given into brackets. This ratio remains stable. There are five times more transition cells than place cells.

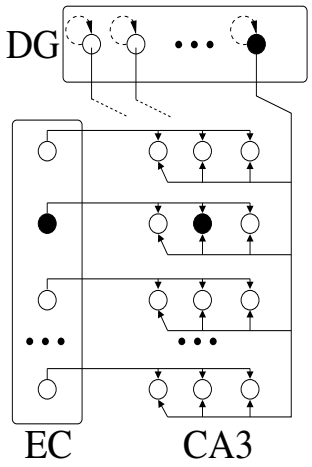


Figure 4: CA3 inputs from EC and DG. For more clarity, only 3 possible transitions are shown (instead of 5 in our simulation). For the same reason, connections from only one neuron of DG are drawn.

Now that we know the number of possible transitions starting from a given place cell, we can modify our model of CA3 (see fig. 4). Instead of having a full matrix linking all location together, we can restrict the possible connections to 5 (as found before). Consequently the number of neurons of CA3 has

decreased since we only take into account real possible transitions and not all the combination of place cells. Each CA3 neurons of a given line receives projections both from EC and DG. Each CA3 neuron belongs to a particular neighboorhood supervised by a single EC neuron (a line in the figure 4). No learning is allowed on those links and their weight are not sufficient to trigger any activity on the associated CA3 neurons. Conversely, each CA3 neuron is connected to all the DG neurons through conditional links. The activation of EC neurons increases the weights coming from the activated neuron in DG. When no CA3 neuron already corresponds to this conjunction. Once those weigths learned, in a prediction mode, the single activity of the corresponding DG neuron allows the activity of the CA3 neuron even if no unconditionnal signal comes from EC.

4 Motor action selection using a neural field

The goal selection mechanism proposes several possible transitions. But how to exploit these informations? We have first experimented strict competition between them, but why do not also use other transitions that contains interesting information about the agent location context (only near transitions, of the current place, are proposed...). So instead of having only one transition win in ACC, we now allow several transitions to be taken into account for the movement.

In our model, basal ganglia and pre-motor cortex are not modelized in details. We rather adopt a functionnal model using a dynamical approach in which the action selection and the motor control are obtained by a stable solution of a dynamical system: the neural field [Amari, 1997]. The properties of the neural field have already been successfully experimented to move the robot's arm by imitation using visual tracking of movement [Andry *et al.*, 2004], or control a robot movement [Schöner *et al.*, 1995; Quoy *et al.*, 2003]. Furthermore, the Neural Field can account for most of the properties (action selection according to contextual inputs, persitence, etc..) exhibited by a neural circuit (the striatum, the globus pallidus (internal and external segment), the subthalamic nucleus, the substantia nigrai (compacta and reticulla)) in more detailed models [K. Gurnett and Redgrave, 2001; K. N. Gurnett and Redgrave, 2001; B. Girard and Prescott, 2002].The limbic loop (connected to the hippocampus via the *core* part of ACC) of the basal ganglia is known to play a role in the motor action, and is considered as the output of our model.. The neural field equation is the following:

$$\tau.\frac{f(x,t)}{dt} = -f(x,t)+I(x,t)+h+\int_{z\in V_x} w(z).f(x-z,t)dz \quad (1)$$

Where $f(x,t)$ is the activity of neuron x, at time t. $I(x,t)$ is the input to the system. h is a negative constant. τ is the relaxation rate of the system. w is the interaction kernel in the neural field activation. A difference of Gaussian (DOG) models these lateral interactions that can be excitatory or inhibitory. V_x is the lateral interaction interval that defines the neighborhood. Without inputs the constant h ensures the stability of the neural field homogeneous pattern since $f(x,t) = h$. In the following, the x dimension will by an angle (direction to follow), 0 corresponding to go straight forward.

The properties of this equation allow the computation of attractors corresponding to fixed points of the dynamics and to local maxima of the neural field activity. Repellors may appear too, depending on the inputs. A stable direction to follow is reached when the system is on any of the attractors.

The angle of a candidate transition is used as input. The intensity of this input depends on the corresponding goal transition activity, but also on its origin place cell recognition activity. If only one transition is proposed, there will be only one input with an angle $x_{targ} = x^*$ and it erects only one attractor $x^* = x_{targ}$ on the neural field. If x_c is the current orientation of the animat, the animat rotation speed will be $w = \dot{x} = F(x_c)$ (see fig. 5, bottom).

Merging of several transition informations depends on the distance between them. Indeed the Amari's equation allows cooperation for coherent inputs associated with spatially separated goals (for us different angles proposed). If the inputs are spatially close, the dynamics give rise to a single attractor corresponding to the average of them (see fig. 5). Otherwise, if we progressively amplify the distance between inputs, a bifurcation point appears for a critical distance, and the previous attractor becomes a repellor and two new attractors emerge. An example of two inputs spatially too far to be merged is described in figure 6.

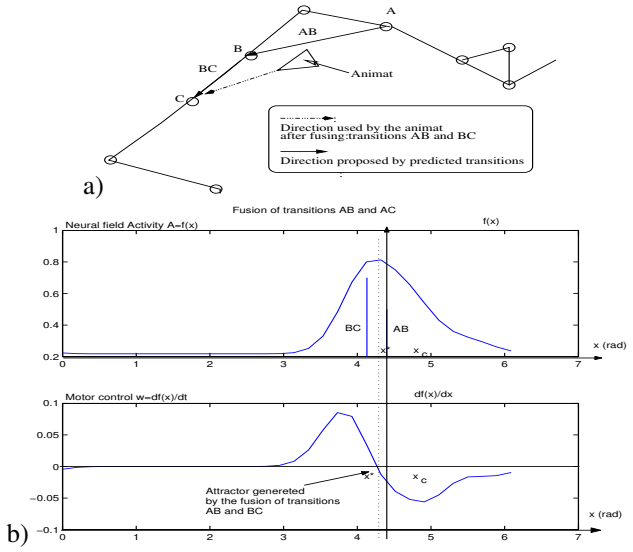


Figure 5: a) Zoom on the cognitive map. The direction followed by the animat corresponds to the attractor generated from both B and BC. b) Activity of the neural field. The inputs of the system (Gaussian) are centered on the directions coded by transitions B and BC. Since the two inputs are spatially close enough, a unique attractor is created (x*).

Oscillations between two possible directions are avoided by the hysteresis property of this input competition/cooperation mechanism. It is possible to adjust this distance to a correct value by calibrating the two elements responsible for this effect: spatial filtering is obtained by convoluting the dirac like signal coming from transition

information with a Gaussian and taking it as the input to the system. This combined with the lateral interactions allows the merging of distinct input as a same attractor. The larger the curve, the more merging there will be.

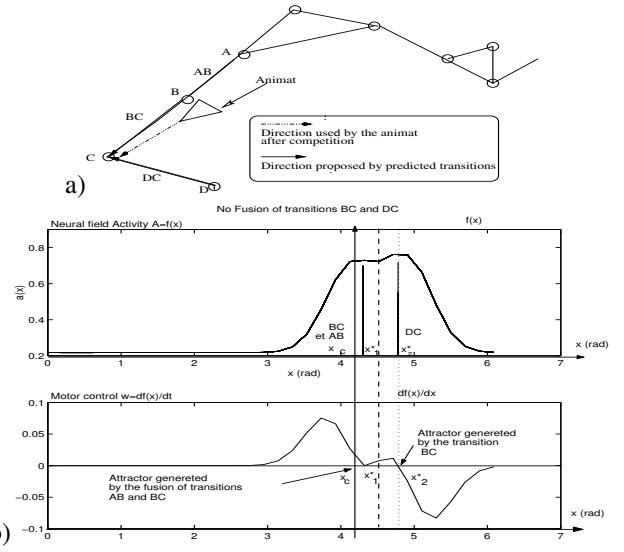


Figure 6: a) Zoom on the cognitive map. Two transitions are predicted: one from place B that gives BC and another from location D that gives DC. Both come to the same place C, but from a different place and so with a different direction. b) Activity of the neural field. Transitions BC and DC are too spatially distant: two attractors (in x_1^* and x_2^*) are created. The motor control converges to x_1^*, closer to the current direction of the animat.

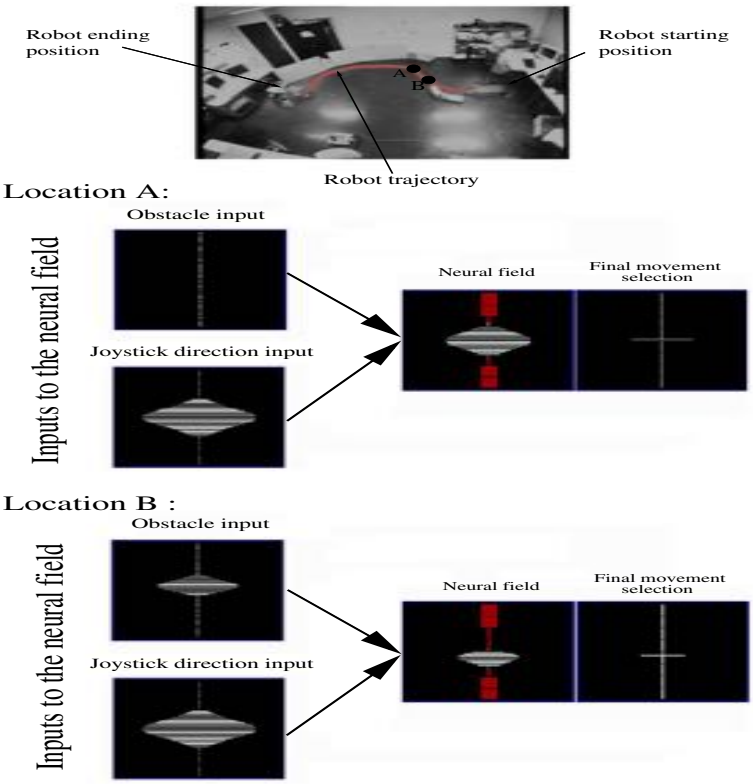


Figure 7: Top: Trajectory of a Labo3 robot in an open environement with obstacles. The direction to go is given by a joystick input. Middle: Neural field activity without any obstacle. The direction taken corresponds to the joystick input. Bottom: Neural field activity with an obstacle. The obstacle shifts the neural field maximal activity leading to a turning move.

In our case, the neural field allows, first to use multiple entries and to combine them to get something coherent, and this at a very low level (motor control and action selection) in the architecture. For example, information from sensors to realize obstacle avoidance can be set as negative inputs generating repellors in the direction leading to an obstacle (see fig. 7). Second, it allows to generalize the decision of the movement and it is robust even in the case of a noisy input signals. This allows to solve the drawback of an incomplete cognitive map and to get a less suboptimal path plan otherwise. Indeed, the inputs give information about where the animat is located inside the place field, since other transitions are only proposed if close to this place (modulation of the transition activity by the recognition). Moreover, the equation takes into account the time information due to a memory effect. Then the effective transition at time 't-1', and so the direction used to enter this place cell, contributes with the informations of other transitions to the dynamics leading to the effective transition at time 't', even if the entry at time 't-1' is not active anymore. This allows to get an effective transition giving a direction to follow with a better accuracy, by taking into account the location of the animat inside the place field of the current place cell (see fig. 8). This memory effect decreases with time and is an important parameter that must be correctly set. It also allows to smooth the different sequences of direction and to perform a motor control and an action selection insensible to its input discontinuities (in particular for stimuli arriving at different times).

Neural fields are a good alternative to overcome the potential fields local minima problems [Khatib, 1986; Koren and Borenstein, 1991]. The main difference is that we do not have a global minimum corresponding to a goal to reach. Hence we cannot draw a global potential landscape (and use a gradient algorithm). In our system there is always a merging between the external information (obstacle) and the *internal* current direction, leading to follow the attractor which is the nearest from the current direction. Indeed, even though there are several attractors due to different obstacles, they are not added like potentials resulting in high "potential barriers" from which the robot cannot escape, or compensating each other. However, we have to tune a parameter in order to get through a door for instance. This parameter is the width of the interaction interval V_x. It depends on the infrared cell responses and on the size of the robot.

5 Conclusions and discussion

Transition based models brings interesting features to action selection problems. They allow to solve several problems in an efficient and very simple way : local minima encountered

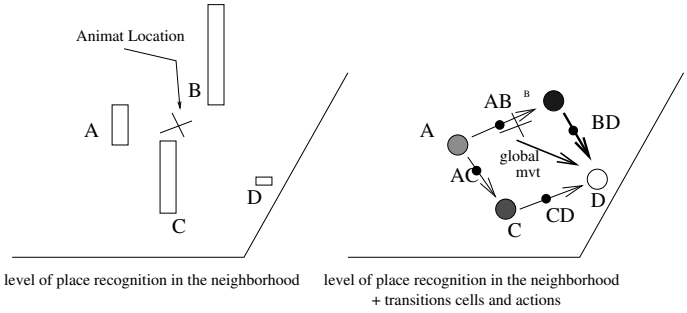


Figure 8: The merging mechanism allows to get a better direction (global mvt) than the use of the single information obtained from current transition (BD). It takes into account the previous movement performed and the transitions predicted from close enough place cell (C).

with steady states models (i.e. : place cells or potential fields) do not appear and the selection of the action is easier to perform without any ambiguity. From a biological point of view, our model avoids an homunculus problem since the selection is done naturally by exploiting the dynamics of the system used. It may be a possible candidate to explain how the hippocampus uses information.

Neural fields allow to merge multi modal inputs to get a stable result. One can imagine to integrate several other signals such as sound direction and other visual target directions. Note that the updating of these different informations does not need to be performed at the same frequency. The neural field is robust enough to deal with intermittent information or high level signal having a very low frequency due to their computation time.

As it has been shown for the motor action selection, relaxing the system constraints can lead to a better generalisation. On going works focus on the generalization of this approach to the transition construction. Indeed we are testing the possibility to allow 'AA' like transitions resulting in the coexistence of place cells and transition cells in CA3. We are also working on the internal hippocampal loop. More exactly, we focus on the link coming from subiculum to the deep layer of EC that might allow to integrate into place cells path integration informations. So integrated movement would allow to activated place cell without visual information... Next, integration of a higher level should allow the animat to discover shortcuts, by integrating the informations of the transitions successively used from the beginning of the planning until reaching the goal. This mechanism should be very useful when the cognitive map is incomplete at the beginning of the exploration, so that the animat could try to use shorter unexperienced paths.

Acknowledgments This work is supported by two french ACI programs. The first one on the modelling of the interactions between hippocampus, prefontal cortex and basal ganglia in collaboration with B. Poucet (CRNC, Marseille) JP. Banquet (INSERM U483) and R. Chalita (LAAS, Toulouse). The second one on the dynamics of biologically plausible neural networks in collaboration with M. Samuelides (SupAro, Toulouse), G. Beslon (INSA, Lyon), S. Thorpe (CERCO, Toulouse) and B. Cessac (INLN, Nice).

References

[Amari, 1997] S. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1997.

[Andry *et al.*, 2004] P. Andry, P. Gaussier, and J. Nadel and. B.Hirsbrunner. Learning invariant sensory-motor behaviors: A developmental approach of imitation mechanisms. *Adaptive behavior*, 12(2):117–140, 2004.

[Arbib and Lieblich, 1977] M.A. Arbib and I. Lieblich. Motivational learning of spatial behavior. In J. Metzler, editor, *Systems Neuroscience*, pages 221–239. Academic Press, 1977.

[Arleo and Gerstner, 2000] A. Arleo and W. Gerstner. Spatial cognition and neuro-mimetic navigation: A model of hippocampal place cell activity. *Biol. Cybern.*, 83(3):287–299, 2000.

[B. Girard and Prescott, 2002] A. Guillot-K. N. Gurnett B. Girard, V. Cuzin and T. J. Prescott. From animals to animats 7. In J. Hallam-G. Hayes B. Hallam, D. Floreano and J. A. Mayer, editors, *Proceedings of the Seventh International Conference on Simulation of Adaptive Behavior*. MIT Press, 2002.

[Bachelder and Waxman, 1994] I. A. Bachelder and A. M. Waxman. Mobile robot visual mapping and localization: A view-based neurocomputationnal architecture that emulates hippocampal place learning. *Neural Networks*, 7:1083–1099, 1994.

[Banquet *et al.*, 1997] J.P. Banquet, P. Gaussier, J.C. Dreher, C. Joulain, and A. Revel. *Cognitive Science Perpectives on Personality and Emotion*, volume 124, chapter Space-Time, Order and Hierarchy in Fronto-Hippocampal System: A Neural Basis of Personality. Elsevier Science BV Amsterdam, 1997.

[Banquet *et al.*, 2005] J. P. Banquet, P. Gaussier, M. Quoy, and A. Revel. A hierarchy of associations in hippocampo-cortical systems: Cognitive maps and navigation strategies. *Neural Computation*, 17:1339–1384, 2005.

[Bellman, 1958] R. E. Bellman. On a routing problem. In *Quaterly of Applied Mathematics*, volume 16, pages 87–90, 1958.

[Bugmann *et al.*, 1995] G. Bugmann, J.G. Taylor, and M.J. Denham. Route finding by neural nets. In J.G. Taylor, editor, *Neural Networks*, pages 217–230, Henley-on-Thames, 1995. Alfred Waller Ltd.

[Donnart and Meyer, 1996] J.Y. Donnart and J.A. Meyer. Learning reactive and planning rules in a motivationnally autonomous animat. *IEEE Transactions on Systems, Man and Cybernetics-Part B*, 26(3):381–395, 1996.

[Gaussier *et al.*, 2000] P. Gaussier, S. Leprêtre, M. Quoy, A. Revel, C. Joulain, and J.P. Banquet. Experiments and models about cognitive map learning for motivated navigation. *Robotics and Intelligent Systems Series*, 24:53–94, 2000.

[Gaussier *et al.*, 2002] P. Gaussier, A. Revel, J.P. Banquet, and V. Babeau. From view cells and place cells to cognitive map learning: processing stages of the hippocampal system. *Biological Cybernetics*, 86:15–28, 2002.

[K. Gurnett and Redgrave, 2001] T. J. Prescott K. Gurnett and P. Redgrave. A computational model of action selection in basal ganglia. i. a new functional anatomy. *Biological Cybernetics*, 84:410, 2001.

[K. N. Gurnett and Redgrave, 2001] T. J. Prescott K. N. Gurnett and P. Redgrave. A computational model of action selection in basal ganglia. ii. analisis and simulation of behavior. new functional anatomy. *Biological Cybernetics*, 84:411–423, 2001.

[Khatib, 1986] 0. Khatib. Real-time obstcle avoidance for manipulators and mobile robots. *Int. Journ. of Rob. Res.*, 5(1):90–98, 1986.

[Koren and Borenstein, 1991] Y. Koren and J. Borenstein. Potential field methods and their inherent limitations for mobile robot navigation. In *Proc. IEEE Conf. on Rob. and Autom.*, pages 1398–1404, 1991.

[Meyer and Wilson, 1991] J. A. Meyer and S. W. Wilson. From animals to animats. In Bardford Books2-4, editor, *First International Conference on Simulation of Adaptive Behavior*. MIT Press, 1991.

[O'Keefe and Nadel, 1978] J. O'Keefe and N. Nadel. *The hyppocampus as a cognitive map*. Clarenton Press, Oxford, 1978.

[Quoy *et al.*, 2003] M. Quoy, S. Moga, and P. Gaussier. Dynamical neural networks for top-down robot control. *IEEE transactions on Man, Systems and Cybernetics, Part A*, 33(4):523–532, 2003.

[Samsonovich and McNaughton, 1997] A. Samsonovich and B. McNaughton. Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience*, 17(15):5900–5920, 1997.

[Schölkopf and Mallot, 1995] B. Schölkopf and H. A. Mallot. View-based cognitive mapping and path-finding. *Adaptive Behavior*, 3:311–348, 1995.

[Schöner *et al.*, 1995] G. Schöner, M. Dose, and C. Engels. Dynamics of behavior: theory and applications for autonomous robot architectures. *Robotics and Autonomous System*, (2–4):213–245, 1995.

[Tolman, 1948] E.C. Tolman. Cognitive maps in rats and men. *The Psychological Review*, 55(4), 1948.

[Trullier *et al.*, 1997] O. Trullier, S. I. Wiener, A. Berthoz, and J. A. Meyer. Biologically based artificial navigation systems: review and prospects. *Progress in Neurobiology*, 51:483–544, 1997.

[Tyrrell, 1993] T. Tyrrell. The use of hierarchies for action selection. *Adaptive Behavior 1*, 4, 1993.

[V. Hok and Poucet, to appear in 2005] P.P. Lenck-Santini V. Hok, E. Save and B. Poucet. Coding for spatial goals in prelimbic-infralimbic area of the rat frontal cortex. *Proceedings of the National Academy of Sciences*, (to appear in 2005).

A computational model of reach decisions in the primate cerebral cortex

Paul Cisek
Université de Montréal
Département de physiologie
C.P. 6128 Succursale Centre-ville, Montréal, QC H3C 3J7, Canada
paul.cisek@umontreal.ca

Abstract

Neurophysiological evidence suggests that visually-guided reaching movements are produced through "specification" and "selection" processes that overlap both temporally and anatomically [Cisek and Kalaska, 2005]. Here, I present a formal computational model which demonstrates how partial specification of several potential movement directions, and the selection of the correct movement, can occur in populations of directionally tuned cells in a distributed cortical network including posterior parietal, premotor, prefrontal, and primary motor cortex. The model reproduces a large set of neurophysiological and psychophysical phenomena, including the behavior of cortical cells during a reach decision task and the spatial and temporal statistics of human reaching choices.

1 Introduction

Traditional theories of voluntary movement control view it as a serial process of planning and execution situated at the end of a larger serial process of perceptual representation and cognitive decision-making. In that view, information from many systems involved in sensory processing, memory, affect, etc. is integrated by cognitive systems to make a decision about the course of action that is appropriate for the current situation. This decision is then turned into a motor program by a motor planning system, and then executed by the appropriate control circuits. However, despite its theoretical appeal, this serial view of the functional architecture of voluntary behavior does not find support in neurophysiological data. In particular, neural systems do not appear to be partitioned into "decision-making", "planning", or "execution" centers but rather appear to mix these putative functions, even at the level of single-cell activity [Cisek, 2005; Kalaska *et al.*, 1998].

As an alternative to these concepts, I present a hypothetical functional architecture that is inspired by neurophysiological data from frontal and parietal cortical regions. The hypothesis is based on a key distinction between processes of "action specification" and "action selection". Action specification involves parallel sensorimotor transformations which use sensory information about the spatial layout of the environment to crudely specify potential actions which are currently available (potential reaching actions, potential grasp points, potential

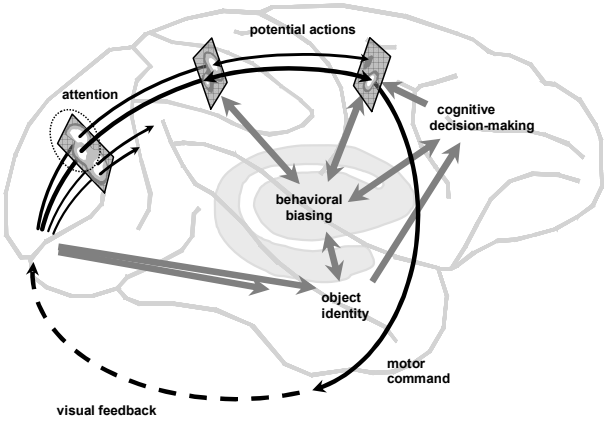


Figure 1: Schematic diagram of the "affordance competition theory" of visually-guided action. Black arrows represent processes which transform visual information into representations of potential actions, and gray arrows represent processes involved in selecting from among these the action that is most appropriate given the current behavioral context. Rectangles represent instances of neural populations which encode information about potential actions as "parameter fields".

gaze targets, etc). Action selection involves processes which narrow down these potential actions, eliminating many from further sensorimotor processing on the basis of salience, attention, behavioral relevance, expected reward, and other cognitive factors. When an action is selected for execution, it is not pre-planned in detail before movement onset, but rather specified crudely and then fine-tuned on-line using rapid feedback through the parietal lobe and predictive feedback through a cerebellar forward model [Miall and Wolpert, 1996].

Figure 1 shows a simplified schematic of how and where specification and selection may take place in the primate brain during visually-guided voluntary behavior [Cisek, 2001]. At all times, information from the dorsal visual stream is used to specify the parameters of several potential motor actions that are currently available (black arrows). These "potential actions" are defined as activity in a series of parameter fields (oblique rectangles), each of which is a neural population in which the activity of a cell indicates the likelihood of performing an action with the parameter values preferred by that cell. Distinct potential actions appear as islands of activity in such fronto-parietal parameter fields. These representations compete for overt execution through mutual inhibition that is biased by various factors such as salience, attention, expected reward, and other cognitive factors, many of which are computed on the basis of information from the ventral stream and processed in the basal ganglia and frontal cortical regions (gray arrows). In brief, it is hypothesized that voluntary behavior involves a constant competition, biased by cognitive influences, among simultaneous early representations of potential actions [Cisek, 2001; Cisek and Turgeon, 1999; Kalaska *et al.*, 1998].

2 Methods

Figure 2a shows the network architecture of a formal computational model which demonstrates how partial specification of several potential movement directions, and the selection of the correct movement, can occur in populations of directionally tuned cells in a distributed cortical network including posterior parietal (PPC), dorsal premotor (PMd), prefrontal (PFC), and primary motor cortex (M1). Briefly, visual information generates activity in a field of tuned PPC neurons, with peaks corresponding to

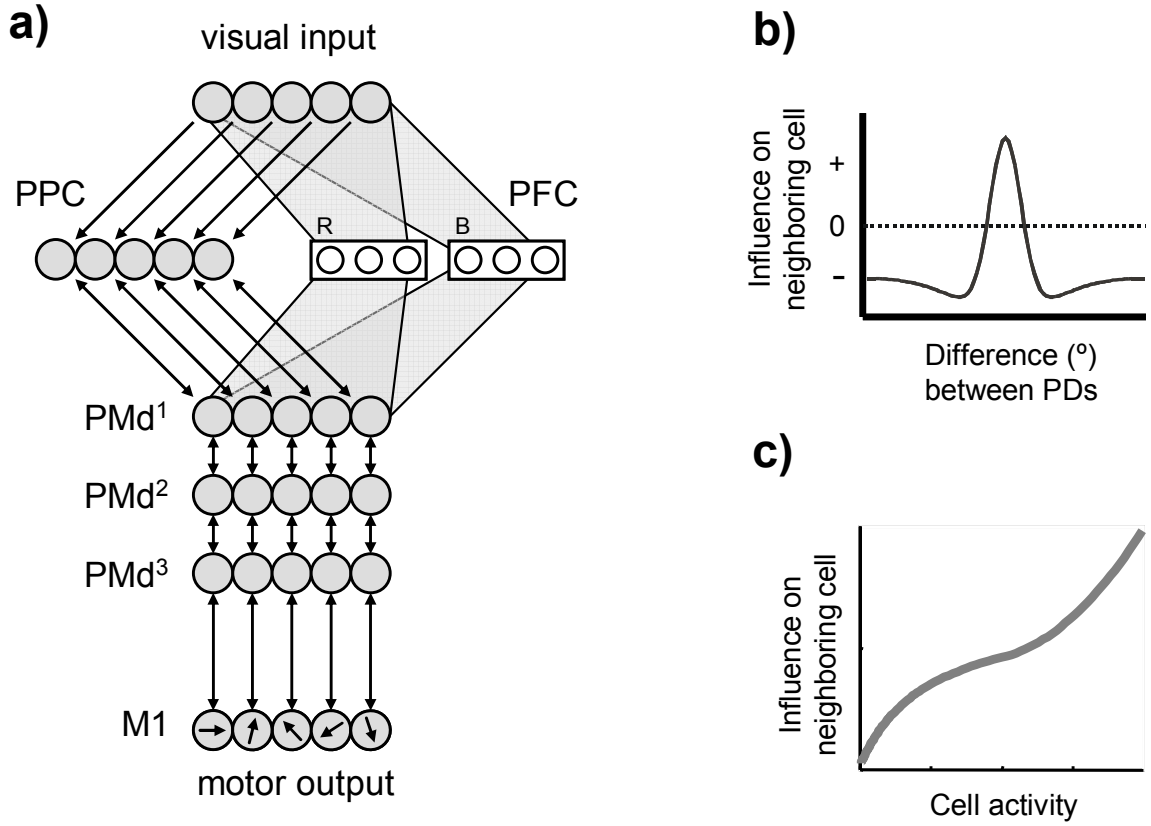


Figure 2: Computational model. A) Network architecture. B) The kernel of interactions between cells with different preferred directions. C) The lateral influence of a cell on its neighbors as a function of activity.

different potential actions. Through reciprocal topological connections, this pattern is repeated in PMd. Because lateral connections among model PPC and PMd cells are organized with an on-center-off-surround pattern, distinct peaks corresponding to distinct actions compete against each other through mutual inhibition. This competition is biased by various factors, notably including excitatory input from model PFC cells which gradually integrate evidence in favor of choosing each of the alternative movement options.

Each layer in the model consists of 90 cells whose behavior is governed by the following non-linear differential equation:

$$\frac{dX}{dt} = -\alpha X + (\beta - X)\gamma \cdot E - X \cdot I + noise$$

where X is the activity of a given neuron, $\frac{dX}{dt}$ is the change in that activity over time, E is the excitatory input to the neuron, I is the inhibitory input, α is a decay rate, β is the neuron's maximum activity, and γ is the excitatory gain. The constants were set at $\alpha = 3$, $\beta = 2$, and $\gamma = 6$ for all populations except the PFC layers, for which they were set at $\alpha = 0.01$, $\beta = 3$, and $\gamma = 0.25$. While the dynamics of all cells in all layers were similar, what gave them different properties was the nature of the excitatory and inhibitory inputs that each received.

There were two kinds of external inputs to the model: 1) visual information about objects in the environment; and 2) a GO signal. Visual information consisted of a vector of binary values indicating the presence or absence of an object of a particular category at a particular location. Processing along the "dorsal stream" was insensitive to category information and cells in the Posterior Parietal Cortex (PPC) layer simply received input whenever any object was present in the direction to which they were tuned. In addition to this excitatory input from "visual" regions, cells within the PPC layer also received excitatory feedback connections from dorsal premotor cortex (PMd), as well as lateral interactions within PPC itself. These interactions involved mutual excitation from neighboring cells with a similar preferred direction (PD) and inhibitory input from cells with different PDs (Fig. 2b). Because of this on-center-off-surround architecture among PPC cells, the visual input was contrast-enhanced, forming distinct broad peaks centered on the directions to given objects. PMd was simulated in a similar manner. That is, it received excitatory input from "upstream" PPC cells as well as feedback excitation from "downstream" regions, and included lateral on-center-off-surround interactions among cells within PMd. Importantly however; the input to PMd from PPC was modulated by biasing signals from the model's prefrontal cortex (PFC).

Because of the positive feedback between PPC and PMd, both regions exhibit similar responses to visual inputs specifying one or more potential targets for movement. For example, if two targets are presented in visual space, distinct peaks of activity will appear, first in PPC and then in PMd, representing the potential actions of moving to each of these targets. Due to lateral interactions within both PPC and PMd, these peaks of activity will compete against each other, but in the absence of information favoring one over the other, their influence will be balanced and both will persist as sustained activity in both PPC and PMd.

The nature of the interaction among cells within individual layers of the model is critical to its behavior. Because neural activities are noisy, the competition between distinct peaks of activity cannot follow a simple "winner-take-all" rule. If that were the case, then random fluctuations due to noise would determine the winner each time, and no informed decision-making would be possible. To prevent this, small differences in the levels of activity associated with two response choices should be treated as uninformative noise and ignored by the system. On the other hand, if the activity associated with one of the choices becomes sufficiently strong due to biasing information in favor of making that choice, then it should be allowed to suppress its opponents and win the competition. In other words, there should be a threshold of activity which, once reached by a particular peak of activity, causes it to be selected as the final response choice. As described by Grossberg [1973], implementing such resistance to noise as well as a decision threshold within a competitive network can be done using a non-linear function defining interactions between neighboring cells. In particular, the function used here is of the form shown in Fig. 2c, with a slower-than-linear portion when activity is low, and a faster-than-linear portion when activity is high. Because of this shape, when two or more peaks are present in the population and have low levels of activity, their influence upon each other de-emphasizes the differences between their activities, thus achieving balance and resistance to noise. However, once the activity of one of the peaks increases and passes into the faster-than-linear regime of the interaction function, then it begins to exert stronger and stronger suppression upon its opponents, thus winning the competition. The point at which a given peak becomes the winner is called a "quenching threshold" [Grossberg, 1973], and it effectively acts as a threshold for committing to a particular decision. However, unlike classical models of decision-thresholds [Carpenter and Williams, 1995; Reddi *et al.*, 2003; Smith and Ratcliff, 2004], the quenching threshold is not a preset constant in the model but rather an emergent threshold which depends both on the number of choices, their relative strengths, and even the angular distance between them.

As described earlier, the competition between distinct regions of activity in PPC and PMd is biased by modulatory input from the prefrontal cortex (PFC). Model PFC cells are sensitive to specific combinations of spatial and category information. Their spatial "receptive fields" are very large but they are topographically connected to similarly tuned PMd cells. Because there are no lateral interactions, and thus no topology, between model PFC cells, they may be thought of as arbitrarily interspersed in the prefrontal cortex.

The dynamics of the PFC cells are slow relative to PPC and PMd (because α and γ are small), and so their activity gradually grows whenever they are presented with inputs that match either their spatial preference or their category preference or both. Thus, they integrate over time the total evidence in favor of a particular choice of action. As that integration proceeds, it biases the competition in PMd until one of the PMd regions of activity crosses the quenching threshold and is thus selected for execution.

The activity in the PMd population is transmitted to the M1 population in the model when the GO signal (a single scalar) is turned on. This begins the selected movement, and further features of execution are not simulated.

3 Results

The network reproduces, with a single set of parameters, a large set of neurophysiological and psychophysical findings. For example, it reproduces the behavior of the kinds of cells observed in PMd during a variety of reach-decision tasks [Cisek and Kalaska, 2005]. Figure 3 compares the activity of the model with neural data collected from PMd and M1 during execution of a "2-target task" in which monkeys were presented with two possible reach choices and then given a cue to choose one for execution. In the simulation, the model was presented with two targets belonging to different categories (red and blue targets). This caused activity to arise in two locations in the PPC layer, as well as in two populations of PFC cells. Due to the positive feedback between PPC and PMd, two peaks of activity were present in both regions, even after the targets themselves vanished. Next, the disambiguating color cue was presented, simulated as excitation to all cells in PFC which preferred the indicated category, i.e. red. This caused a bias to arise in PFC and to tip the balance in PMd in favor of the target where a red cue had appeared, pushing the corresponding activity over the quenching threshold and thus toward a decision to move to that target. In addition to reproducing this main result, the model also exhibited more subtle phenomena such as an inverse relationship between the number of targets and the magnitude and width of activity associated with each, and the gradient of properties between rostral and caudal PMd subpopulations, which were implemented with identical equations. The model also reproduced the observation that decision errors are in most

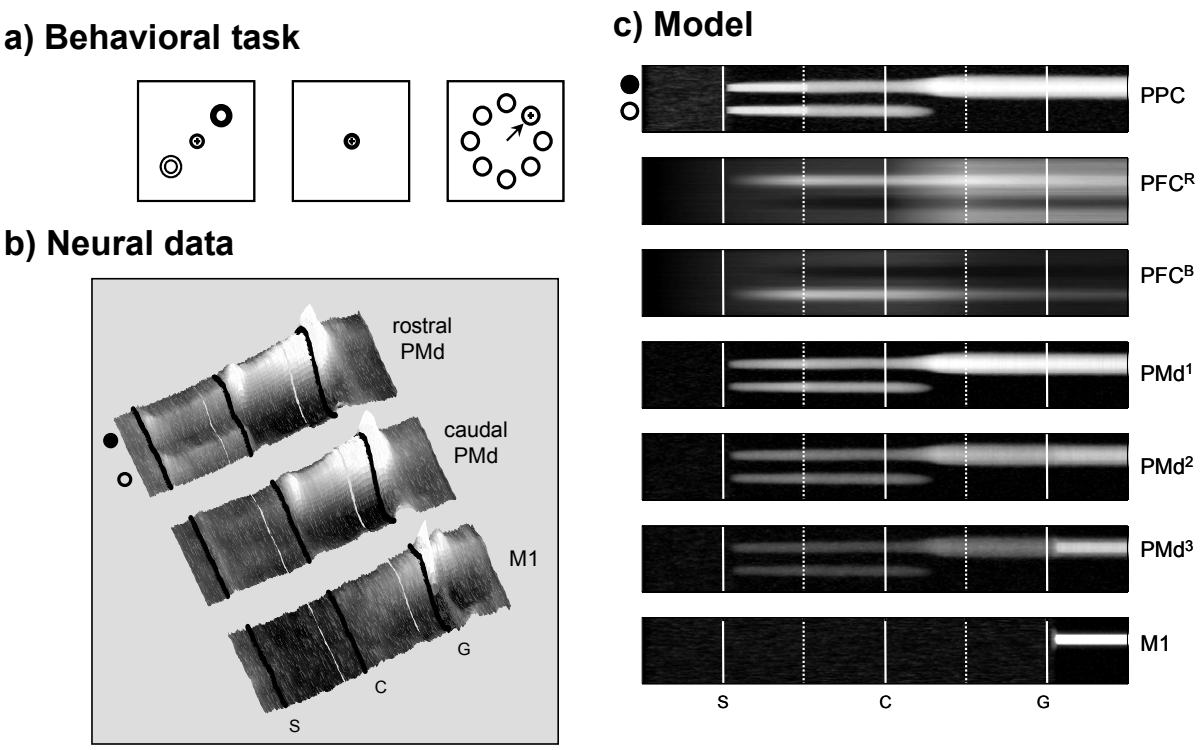


Figure 3: Comparison of model and data. A) The behavioral task involved first presenting two potential reach targets, then indicating which was the correct target, and finally presenting a GO instruction. B) Population activity in rostral PMd, caudal PMd, and M1, represented as greyscale plots where time is indicated along the x-axis and cell preferred direction along the y-axis. The activity in PMd indicated the presence of two directional signals after presentation of the spatial cues (S), the selection of one of these after the disambiguating color cue (C), and execution of the selected target after the GO signal (G). C) Activities in the model populations during simulation of an analogous task. Comparison of the PMd and M1 activities between the model and data shows similarity among most of the salient qualitative properties. Neural data adapted from [Cisek and Kalaska, 2005].

cases caused by a bias in the activity which existed *prior* to the choice cue (not shown).

The model also simulates several psychophysical results on the spatial and temporal statistics of human reaching choices. For example, Ghez et al. [1997] reported that when choices are made quickly, subjects move in-between targets that are close together, but choose randomly between targets that are far apart. The same pattern of errors is produced by the model when the time between the choice cue and the GO signal is made sufficiently short. As shown in Fig. 4a, when targets are far apart, the decision is forced by the strong winner-take-all dynamics in M1, and the choice is determined by noise. In contrast, when targets are close together the two peaks of activity in PMd coalesce and the resulting movement is in-between the two targets.

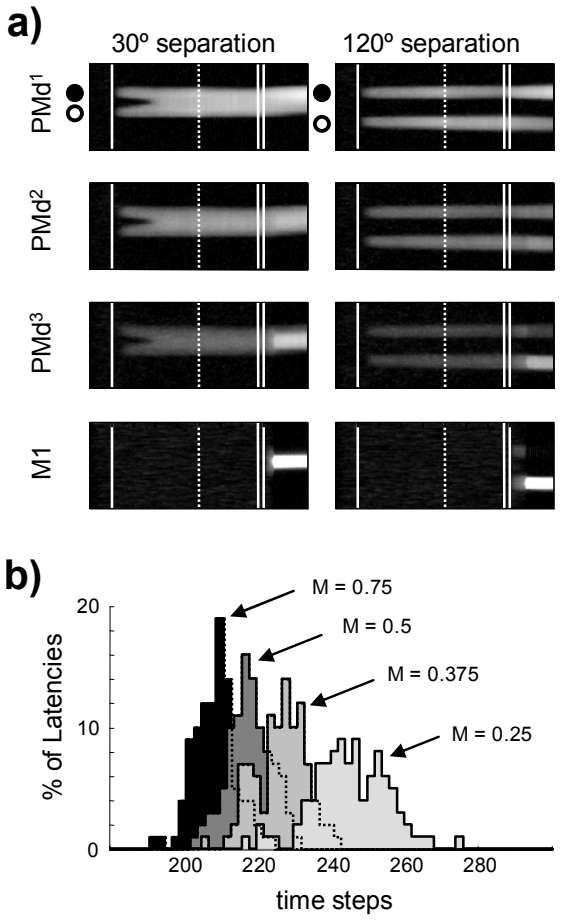


Figure 4: A) Simulations of the Ghez paradigm, in which humans make choices only a short time after choice cue presentation. Left: When targets are close together, the peaks of activity corresponding to each coalesce, and the resulting movement is in-between the targets. Right: When targets are far apart, selection occurs but is determined by noise. In the case shown, the wrong target was selected. B) Distributions of the decision latency as a function of choice cue magnitude (M).

The model also produces results on the distribution of decision-latencies. Because of the gradual accumulation of activity in PFC until the quenching threshold is reached, the behavior of the network resembles that of horse-race models of decision making, and like those models, reproduces results on the distribution of reaction times with different levels of decision certainty [Carpenter and Williams, 1995; Reddi *et al.*, 2003; Smith and Ratcliff, 2004]. Figure 4b shows results of simulations of a 2-target task with four different levels of magnitude of the choice cue. As shown for human data, as the quality of the information provided by the choice cue decreases, distributions of decision latencies become both later and broader.

4 Conclusions

The model presented here is a mathematical implementation of a general hypothesis on action selection and motor planning – that in many everyday situations, several potential actions are often specified simultaneously and compete for overt execution. According to this view, visual information in the dorsal stream [Milner and Goodale, 1995; Ungerleider and Mishkin, 1982] is used to specify the parameters of potential motor actions in parietal [Kalaska *et al.*, 1998; Snyder *et al.*, 1997] and premotor cortex [Cisek and Kalaska, 2005]. These potential actions compete through lateral inhibition, while at the same time other systems collect evidence for or against particular actions, in part using information from the ventral stream. These mechanisms include selective attention [Allport, 1987; Neumann, 1990; Tipper *et al.*, 1998], switching mechanisms in the basal ganglia [Redgrave *et al.*, 1999], and accumulation of task-relevant information in prefrontal cortex [Hoshi *et al.*, 2000; Kim and Shadlen, 1999]. This hypothesis is an attempt to unify psychophysical results on human decision-making with data on single cell activity in monkeys during simple decision tasks, and to understand from a theoretical perspective the functional reason for the observed mixing of sensory, motor, and cognitive variables within the activity of cells in movement-related cortical regions.

5 References

[Allport,D.A., 1987] Allport,D.A. Selection for action: Some behavioral and neurophysiological considerations of attention and action. Heuer, Herbert and Sanders, Andries F. *Perspectives on Perception and Action*. (15), 395-419, Hillsdale, NJ, Lawrence Erlbaum Associates. 1987.

[Carpenter,R.H. and Williams,M.L., 1995] Carpenter,R.H. and Williams,M.L. Neural computation of log likelihood in control of saccadic eye movements. *Nature* 377(6544), 59-62, 1995.

[Cisek,P., 2001] Cisek,P. Embodiment is all in the head. *Behavioral and Brain Sciences* 24(1), 36-38, 2001.

[Cisek,P., 2005] Cisek,P. Neural representations of motor plans, desired trajectories, and controlled objects. *Cognitive Processing* 6, 15-24, 2005.

[Cisek,P. and Kalaska,J.F., 2005] Cisek,P. and Kalaska,J.F. Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. *Neuron* 45(5), 801-814, 2005.

[Cisek,P. and Turgeon,M., 1999] Cisek,P. and Turgeon,M. 'Binding through the fovea', a tale of perception in the service of action. *Psyche* 5(34), 1999.

[Ghez,C., Favilla,M., Ghilardi,M.F., Gordon,J., Bermejo,R., and Pullman,S., 1997] Ghez,C., Favilla,M., Ghilardi,M.F., Gordon,J., Bermejo,R., and Pullman,S. Discrete and continuous planning of hand movements and isometric force trajectories. *Experimental Brain Research* 115(2), 217-233, 1997.

[Grossberg,S., 1973] Grossberg,S. Contour enhancement, short term memory, and constancies in reverberating neural networks. *Studies in Applied Mathematics* 52, 213-257, 1973.

[Hoshi,E., Shima,K., and Tanji,J., 2000] Hoshi,E., Shima,K., and Tanji,J. Neuronal activity in the primate prefrontal cortex in the process of motor selection based on two behavioral rules. *Journal of Neurophysiology* 83(4), 2355-2373, 2000.

[Kalaska,J.F., Sergio,L.E., and Cisek,P., 1998] Kalaska,J.F., Sergio,L.E., and Cisek,P. Cortical control of whole-arm motor tasks. Glickstein, M. *Sensory Guidance of Movement, Novartis Foundation Symposium #218*. 176-201, Chichester, UK, John Wiley & Sons. 1998.

[Kim,J.-N. and Shadlen,M.N., 1999] Kim,J.-N. and Shadlen,M.N. Neural correlates of a decision in the dorsolateral prefrontal cortex of the macaque. *Nature Neuroscience* 2(2), 176-185, 1999.

[Miall,R.C. and Wolpert,D.M., 1996] Miall,R.C. and Wolpert,D.M. Forward models for physiological motor control. *Neural Networks* 9(8), 1265-1279, 1996.

[Milner,A.D. and Goodale,M.A., 1995] Milner,A.D. and Goodale,M.A. The Visual Brain in Action. Oxford University Press. 1995.

[Neumann,O., 1990] Neumann,O. Visual attention and action. Neumann, Odmar and Prinz, Wolfgang. *Relationships Between Perception and Action: Current Approaches*. 227-267, Berlin, Springer-Verlag. 1990.

[Reddi,B.A.J., Asrress,K.N., and Carpenter,R.H.S., 2003] Reddi,B.A.J., Asrress,K.N., and Carpenter,R.H.S. Accuracy, information, and response time in a saccadic decision task. *Journal of Neurophysiology* 90(5), 3538-3546, 2003.

[Redgrave,P., Prescott,T.J., and Gurney,K., 1999] Redgrave,P., Prescott,T.J., and Gurney,K. The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience* 89(4), 1009-1023, 1999.

[Smith,P.L. and Ratcliff,R., 2004] Smith,P.L. and Ratcliff,R. Psychology and neurobiology of simple decisions. *Trends Neurosci.* 27(3), 161-168, 2004.

[Snyder,L.H., Batista,A.P., and Andersen,R.A., 1997] Snyder,L.H., Batista,A.P., and Andersen,R.A. Coding of intention in the posterior parietal cortex. *Nature* 386, 167-170, 1997.

[Tipper,S.P., Howard,L.A., and Houghton,G., 1998] Tipper,S.P., Howard,L.A., and Houghton,G. Action-based mechanisms of attention. *Phil.Trans.R.Soc.Lond.B* 353(1373), 1385-1393, 1998.

[Ungerleider,L.G. and Mishkin,M., 1982] Ungerleider,L.G. and Mishkin,M. Two cortical visual systems. Ingle, D. J., Goodale, Melvyn A., and Mansfield, R. J. W. *Analysis of Visual Behavior*. (18), 549-586, Cambridge, MA, MIT Press. 1982.

Recognizing Invisible Actions

James Bonaiuto, Edina Rosta, and Michael Arbib
Computer Science, Chemistry, Neuroscience, USC Brain Project
University of Southern California
bonaiuto@usc.edu, rosta@almaak.usc.edu, arbib@pollux.usc.edu

Abstract

Action selection in social situations often depends on recognizing the action of another individual, and that individual's intent in executing that action. A class of neurons has been identified in area F5 of the macaque premotor cortex that respond to the observation of both self-directed actions and similar actions performed by others. It is thought that these "mirror neurons" provide the monkey with a common representation for action generation and perception, mediating action recognition and possibly action selection via response facilitation or affordance learning. This system was previously modeled as a feedforward network of artificial neurons that was trained using back propagation to classify several types of grasps based on longer and longer prefixes of a trajectory relating view of hand and object. Since the coding required for the input to the feed-forward model is unrealistic, we present here an updated model of the grasp-related mirror neuron system that utilizes a recurrent neural network trained using back propagation through time and includes audio input, a working memory, and dynamic remapping. The model not only replicates the findings of the original study but also explains further experimental data showing that F5 mirror neurons in the macaque respond to audio as well as visual stimuli and also to visual observation grasps even when the end of the trajectory is hidden and must be inferred.

1 Introduction

Action selection involves formulating a plan to execute the most appropriate action given the state of the environment. In social organisms the environment often includes other agents carrying out their own plans of action. To function effectively in such an environment, the activities of these agents must be recognized and used to guide subsequent action. The relationship between action recognition and action generation has been investigated in the context of cooperative teams of robots [Parker, 1995] and skill learning through imitation [Schaal, 1999]. These two complementary processes are thought to be linked in primate brains by the premotor cortex where it seems that an observed action is mapped onto the internal motor programs used to generate a similar action [Arbib, 2005; Gallese *et al.*, 1996; Rizzolatti *et al.*, 1996].

A class of neurons has been identified in area F5 of the macaque premotor cortex that respond to the observation of both self-directed actions and similar actions performed by others [Gallese *et al.*, 1996]. It is thought that these "mirror neurons" provide the monkey with a common representation for action generation and perception, mediating action recognition and possibly action selection via response facilitation or affordance learning. Moreover, it has been found that F5 mirror neurons in the macaque respond to audio as well as visual stimuli [Koller *et al.*, 2002] and actions where the final component is hidden and must be inferred [Umilta *et al.*, 2001]. Here, we present an updated model of the grasp-related mirror neuron system that utilizes a recurrent neural network trained using back propagation through time and includes audio input, a working memory, and dynamic remapping to address these data.

A previous model [Oztop & Arbib, 2002] of the monkey mirror system used a feed-forward artificial neural network with one hidden layer. Hand state information for a grasp was represented as a 7 dimensional trajectory encoding hand-object relations. At each point in time, the initial segment of the hand state trajectory up to that time was fitted by a cubic spline, and then sampled 30 times to produce a 210 dimensional input vector to the network. In this way, the temporal representation of hand state was pre-processed such that it could be encoded in a spatial representation for input into the feed-forward network. The network was trained on a set of self-performed grasps using back propagation. This system was shown to correctly classify observed grasps, often before the hand contacted the object. In addition, the network yielded neurophysiological predictions concerning the time course of mirror neuron activation and activity during the resolution of an ambiguous grasp.

However, the unnatural temporal to spatial encoding transformation required for the hand-state trajectory (relating hand and object) led us to look for a model that could process the time series of hand-object relationships without extensive recoding. We thus turned to recurrent networks. These networks have been shown to be computationally powerful and useful for reducing problem dimensionality [Jones, 1992].

They have been applied to a variety of problems including context-sensitive language processing [Steijvers & Grunwald, 1996], noisy time series prediction [Giles *et al*., 2001], and temporal sequence classification. We investigated the use of a Jordan-type recurrent network to classify grasps based on the temporal sequence of hand state information. The recurrent network allows us to avoid the input pre-processing steps necessary to recognize sequences with a feed-forward net. The raw 7 dimensional hand state vector is simply input to the network at each time step. The network was again trained on a set of self-generated grasps, but this time using back propagation through time [Werbos, 1990]. We show that this system also correctly classifies different types of grasps, often before the feed-forward implementation does, and preserves the neurophysiological predictions made by the model. Moreover, we have extended the model in a fashion consistent with available data on the macaque brain to explain a range of further experimental data.

Natural actions typically involve both a visual and an audio component. The audio properties of mirror neurons are of major interest because they may have been crucial in the transition from gesture to vocal articulation in the evolution of language [Arbib, 2005]. [Koller *et al*., 2002] (see Figure 5, left) found that a portion of the mirror neurons in area F5 of the macaque premotor cortex that are responsive for the observation of noisy actions (such as peanut breaking and paper ripping) are also just as responsive for the sounds of these actions. Area F5 is located in the ventrorostal portion of area 6 in the caudal inferior arcuate sulcus [Rizzolatti *et al*., 1996]. Audio information reaches inferior caudal arcuate cortex via direct connections from the auditory cortex [Deacon, 1992] and reaches area 6 via indirect connections through area 8 [Arikuni *et al*., 1988; Romanski *et al*., 1998]. The macaque nonprimary auditory cortex has been found to respond to complex sounds [Rauschecker *et al*., 1995] while the primary auditory cortex was found to have a tonotopic organization [Morel *et al*., 1993]. It thus seems that auditory input is extensively preprocessed by the time it reaches premotor cortex.

We model this sound pre-processing by auditory cortex at an abstract level with a non-neural auditory cortex implementation. This abstract model of auditory cortex associates a given type of sound with an arbitrary and distinct pattern of activity in a 9-dimensional vector. This vector is then applied to audio input units that are directly and fully connected to the output layer of the recurrent neural network (figure 2), corresponding to a direct connection from auditory cortex to F5. These connection weights are modified using Hebbian learning. This makes the audio input units and their connections to the external output units behave as a somewhat independent network. In this way, any sound that is consistently perceived during the course of an executed action becomes associated with that action and incorporated into its representation. This type of audio information is inherently actor-invariant and this allows the monkey to recognize that another individual is performing that action when the associated sound is heard. In all actions tested by [Koller *et al*., 2002] the sound was associated with the final phase of the action. Mirror neurons responsive to audio and visual stimuli were thus found to be

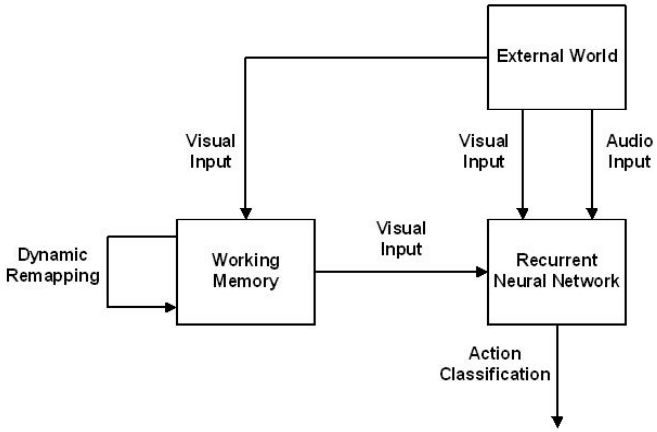


Figure 1: System diagram. The recurrent network receives both visual and auditory input from the external world. Visual input is also input into working memory. When visual information is not available externally, the working memory trace is input into the recurrent network.

active later during audio only conditions than conditions with a visual component. The activation of these neurons during conditions with only audio information was confined to the duration of the audio input.

[Umilta *et al*., 2001] (see Figure 5 left) have shown that mirror neurons in the macaque monkey can infer the result of an action whose final portion is hidden. In these experiments, the monkeys were shown an object that was then obscured by a screen. When the monkey observed the experimenter reaching behind the screen to grasp the object, the same mirror neurons were activated that responded to a visible grasp to the same object. The same neuron does not respond to a reach when no object is visible, or if the human reaches behind a screen that was not previously shown to conceal an object. To recognize that another individual is executing a grasping action even when the goal and final component of that action is hidden, the monkey must possess a working memory trace of the object that the action is directed towards. It is not clear whether or not a working memory representation of the hand is used to extrapolate the grasp trajectory or if the initial hand state trajectory coupled with object location working memory is sufficient to correctly activate F5 mirror neurons. This could be determined by gradually receding the point in the grasp at which the experimenters hand disappears behind the screen until the hand is behind the screen for entire grasp duration. If the mirror neuron activity decays accordingly, this could be evidence that a transient working memory activation is supplying hand state information to area F5 when the hand is not visible.

Dynamic remapping is a process where perceptual representations are updated based on generated motor commands, or related perceptual information. It has previously been used in a model of saccade generation [Dominey & Arbib, 1992] to update the working memory representation of the position of a secondary saccade target based on self-generated eye movements to a primary saccade target. We use dynamic

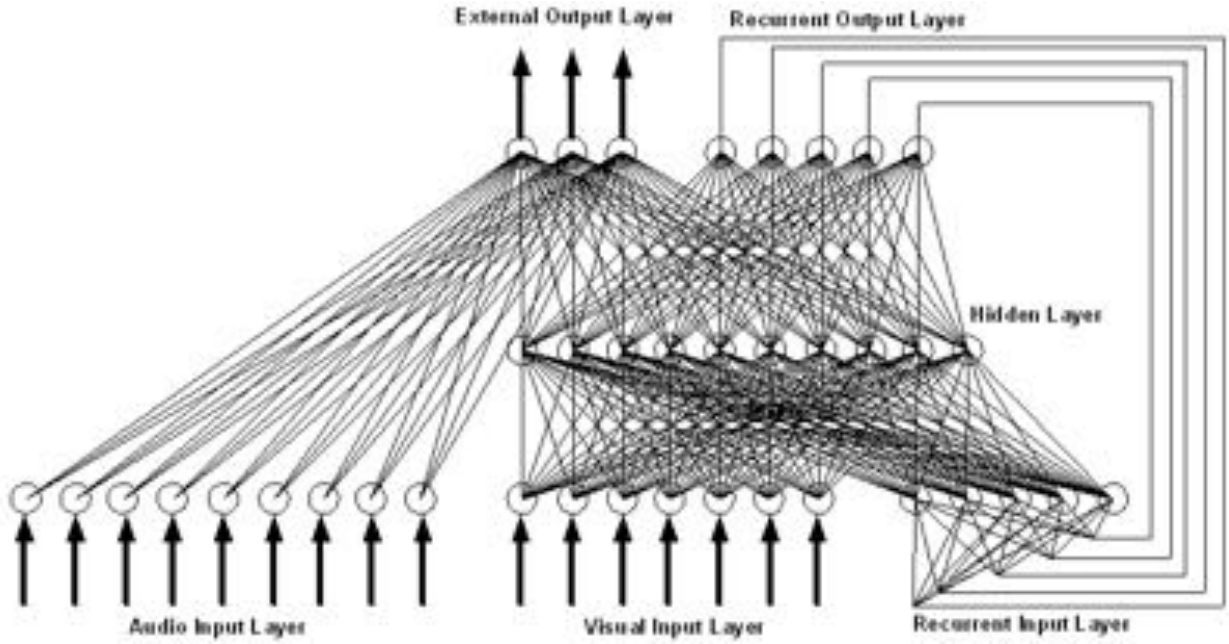


Figure 2: Recurrent network diagram. The visual and recurrent input layers are fully connected to the hidden layer, which is fully connected to the external and recurrent output layers. The recurrent output layer is fully connected to the recurrent input layer and the audio input layer is fully connected to the external output layer.

remapping to extrapolate the observed grasp trajectory once the hand disappears behind the screen. At each point in time that the hand is obscured by the screen, the movement of the still-visible elbow is used to update the working memory representation of the wrist position. To test whether or not this process is actually employed by the primate mirror system, a fake hidden grasp where the hand overshoots the object could be presented to the monkey. If the grasp-related mirror neurons still respond to this example of a hidden non-grasp, this could be an indication that hand trajectory extrapolation is not used in hidden grasp recognition.

The mechanisms of working memory and dynamic remapping of the representations held in working memory allow the model to recognize grasps even when the final stage of object contact is hidden and must be inferred. Before being hidden, the object position and its affordance information are stored in working memory. Once the hand is no longer visible, the working memory representation of wrist position is updated by using the still-visible elbow position and the fact that the forearm is a rigid body. In this way, if the model observes an object which is then hidden by a screen, and then observes a grasp that disappears behind that screen, the wrist trajectory will be extrapolated to end at the remembered object location and the grasp will be recognized.

2 Methods

2.1 Reach and Grasp

We used the multi-joint 3D kinematics simulator developed in [Oztop & Arbib, 2002] to plan a grasp and reach trajectory and execute it in a simulated 3D world. This simulator is a non-neural implementation of the FARS model of primate grasping [Fagg & Arbib, 1998] that controls a virtual 19 degrees of freedom (DOF) arm/hand and performs realistic grasps. Grasps are planned by determining the points of desired contact of fingers on the object (based on the type of grasp: power, precision, or side) and then finding the required arm/hand joint configuration to produce this grasp (the inverse kinematics problem). The final arm/hand joint configuration is determined by gradient descent with noise and the grasp trajectory is then generated by warping time with a cubic spline. The parameters of this spline are derived from empirical studies to fit the natural reach-to-grasp aperture and velocity profile. This simulator was used to generate realistic grasps to train and test the model.

2.2 Visual Analysis of Hand State

The visual information input into network for grasp recognition is the trajectory of a 7-dimensional vector encoding hand-object relations (the hand state). This information is calculated from the joint configuration of the simulated arm/hand and 3D object. The components of the hand state are a(t): aperture of virtual fingers involved in grasping, o1(t): angle between the object axis and the (index finger tip - thumb tip) vector, o2(t): angle between the object axis and the (index finger knuckle - thumb tip) vector, o3(t), o4(t): angle between the thumb and the side of the hand, and the thumb and the inner surface of the palm, d(t): distance to target at time t, and v(t): tangential velocity of the wrist. The hand state is calculated in an object-centered framework, allowing self-generated and observed grasps to evoke similar hand state trajectories. At each point in time that the hand and object are

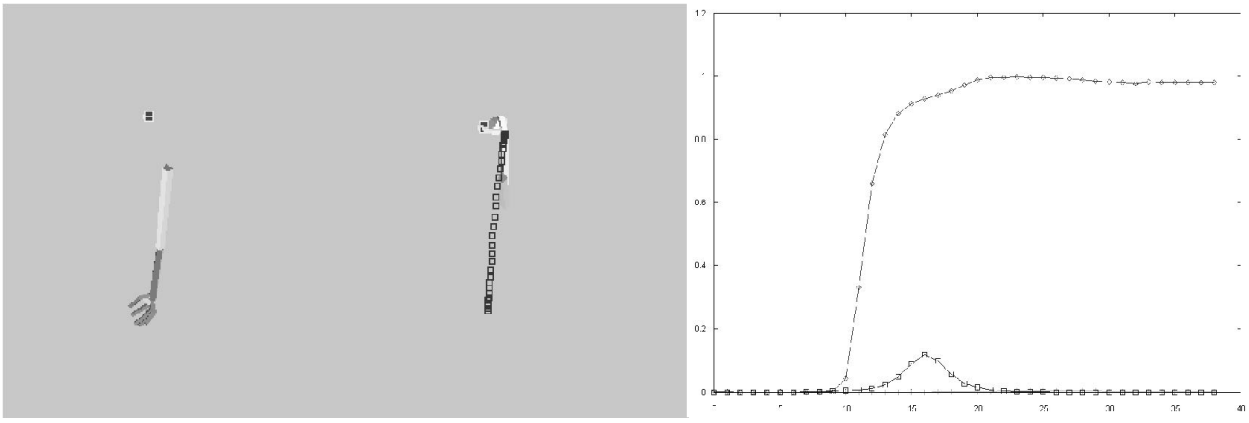


Figure 3: Left: Example of a generated precision grasp. The squares denote the wrist trajectory. Right: External output unit activation for this grasp in figure with no audio input. The precision grasp is correctly recognized well before the hand contacts the object. In all output unit activation figures, the line with the square data points represents the trajectory of the power grasp output unit's activation, and the line with the diamond data points represents the activation of the precision grasp output unit.

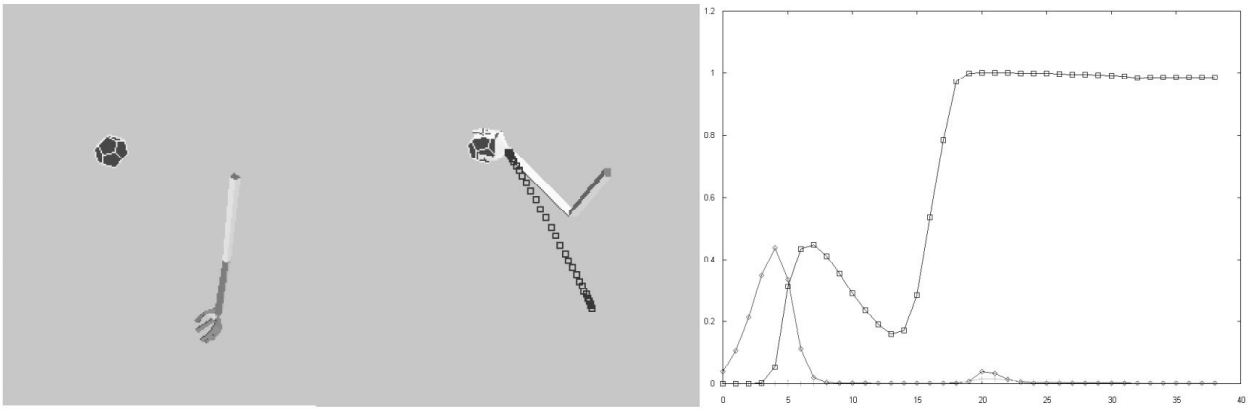


Figure 4: Left: Example of a generated power grasp. The squares show the wrist trajectory. Right: External output unit activation for this grasp with no audio input. The initially ambiguous grasp is correctly recognized as a power grasp well before the hand contacts the object.

visible, this hand state is calculated directly from the simulated arm/hand and object and applied to the input layer of the recurrent neural network. The object and wrist coordinates are stored in working memory. At each point in time that the hand or the object is invisible, the hand state is calculated from the working memory and input to the recurrent neural network.

2.3 Action Recognition

Grasps are recognized by audio and visual input into a recurrent neural network augmented with working memory. The visual input is the 7 dimensional hand state, and the auditory input is an arbitrary, but unique 9 dimensional vector distinguishing different actions. The network output is a 3 dimensional vector, each element of which encodes a type of grasp (power, precision, or side). The most active element in the network's output layer indicates the grasp classification.

Recurrent Network Setup

We used a Jordan-type recurrent network containing 7 external input units, 5 recurrent input units, 10 hidden units, 3 external output units, and 5 recurrent output units. Each layer is fully connected with the layer above it, and the recurrent output units are fully connected with the recurrent input units (see Figure 2). The learning algorithm used is back propagation through time (BPTT) [Werbos, 1990].

Audio Input

Each type of grasp (power, precision, and side) was associated with a unique sound. These sounds are assumed to be pre-processed by auditory cortex (not modeled here) and so in this model are simply represented by arbitrary patterns of audio input activity that are distinct for each type of sound. Audio input to the model was as an array of 9 external input units directly and fully connected with the external output layer of the recurrent neural network. During the last time steps of each type of grasp, a unique pattern of activation of

is presented to the auditory input units. These patterns are of different duration for each grasp type. Activity in these units is propagated to the external output layer along with the hidden layer activity (see Figure 2). For the audio input to external output connection weights, learning is Hebbian if the external output unit is greater than half active, and anti-Hebbian otherwise.

$$W_{ij} = \begin{cases} W_{ij} + \eta A_j(t) MR_i(t), & if MR_i(t) > 0.5 \\ W_{ij} - \eta A_j(t) MR_i(t), & if MR_i(t) \leq 0.5 \end{cases} \quad (1)$$

where $A_j(t)$ is the activity of the jth audio input unit and $MR_i(t)$ is the activity of the ith external output unit at time t, W_{ij} is the weight of the connection between A_j and MR_i, and η is the Hebbian learning rate. The value used in our simulations for η was 0.01. These connections were then normalized using the sum of connection weights to each output unit divided by 5.0. The division of this sum by the constant 5.0 serves as a scaling factor so that normalization bounds connection weights by 0.0 and 5.0.

$$W_{ij} = W_{ij} / ((\sum_i W_{ij})/5.0) \quad (2)$$

Working Memory and Dynamic Remapping

Working memory was implemented as arrays holding 3D coordinates for both the object and the hand. For each time step that the object or hand were visible, their coordinates were stored in their respective working memory array. For each time step that either the hand or object was not visible, random values between -0.5 and 0.5 were added to each x,y,z coordinate value in their working memory array to simulate memory decay. The values held in working memory were used to compute the hand state for network input when either the hand or object was invisible.

Dynamic remapping was carried out on the working memory representation of the wrist position in each time step that the hand was not visible. This serves to update the wrist position in working memory by extrapolating its trajectory. The wrist position working memory was displaced by the same magnitude and direction as the change in elbow position from the previous step in a similar manner to the dynamic remapping of saccade target location based on eye position in [Dominey & Arbib, 1992]. This is accomplished by calculating the difference in elbow position between the two latest time steps, and using this value to update the working memory representation of wrist position:

$$WM_{wrist} = WM_{wrist} + (elbow(t) - elbow(t-1)) \quad (3)$$

where WM_{wrist} is the working memory representation of the wrist's 3 dimensional coordinates and $elbow(t)$ is the three dimensional coordinates of the elbow at time t.

2.4 Training

The training set was constructed by making the simulator perform various grasps in the following way.

1. The objects used were a cube of changing size (a generic size cube scaled by a random factor between 0.5 and 1.5) and a ball (approximated as a dodecahedron, again scaled randomly by a number between 0.75 and 1.5). In the training set formation, a certain object always received a certain grasp and each type of grasp is associated with a distinct audio input pattern at the grasp completion.

2. The object locations were chosen from a portion of the surface of a sphere centered on the simulated arm's shoulder joint. The portion was defined by bounding the longitude and latitude lines on the sphere's surface by -45^o and 45^o. During the generation of training data, this portion of the sphere's surface was traversed in increments of 10^o. Thus the simulator made $9 \times 9 = 81$ grasps per object. Unsuccessful grasp attempts were identified as those in which the resulting trajectory did not bring the hand in contact with the object and were discarded from the training set. For each successful grasp, one negative example was added to the training set to stress that the distance to target was important.

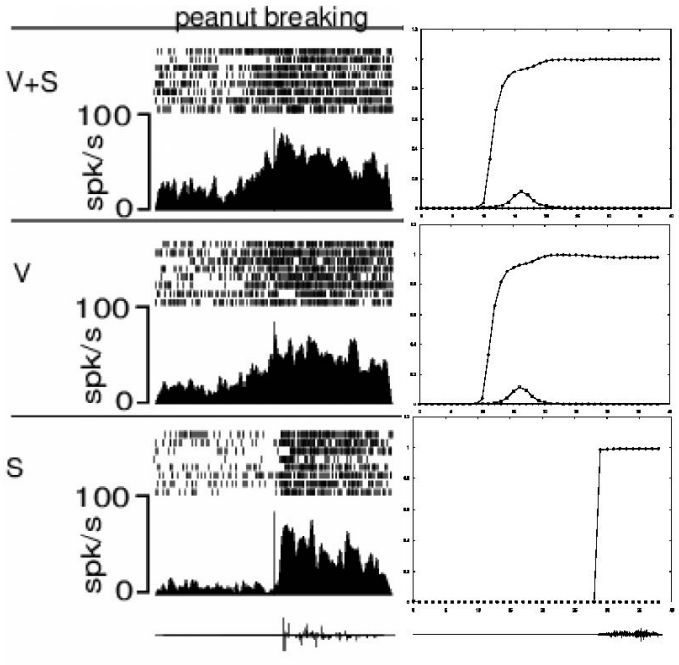


Figure 5: Left: Activation from (Kohler et al., 2002) of an audiovisual mirror neuron responding to (from top to bottom) the visual and audio components, visual component alone, and audio component alone of a peanut-breaking action. At the bottom is an oscillogram of the peanut breaking sound. Right: Activation of the model's external output layer when presented with a precision grasp sequence containing (from top to bottom) visual and audio, visual only, and audio only information. The unit encoding the precision grasp shows the greatest level of activation, while the unit corresponding to power grasps shows a small level of transient activitiy. At the bottom is an oscillogram of the sound associated with the precision grasp. The experimental data and model output show anticipatory mirror neuron activity for visual only and audiovisual conditions, but this activity is confined to the duration of the action sound in the audio only condition.

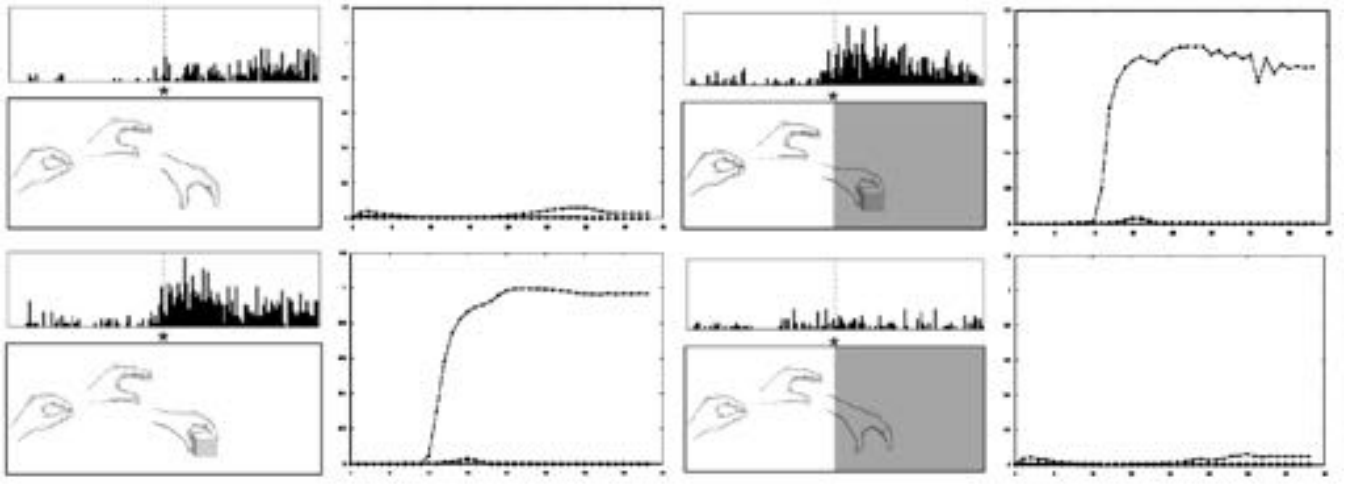


Figure 6: Left: Mirror neuron activation in Umilta et al. 2001 for visible pantomimed grasp, visible grasp, hidden grasp, and hidden pantomimed grasp conditions. All grasps were power grasps. Right: Activation of the model's external output units to these conditions. The only output unit showing a significant level of activity in any plot is the one encoding power grasps. The model output is in good agreement with the experimental data in that visible and hidden grasps are correctly identified, while visible and hidden pantomimed grasps elicit little or no response.

The target location was perturbed and the grasp was repeated (to the original target position).

The recurrent neural network was trained with this set using back propagation through time.

3 Simulation Results

3.1 Recurrent neural network performance

After training, the recurrent network was able to efficiently classify grasps given the hand state trajectory. Most grasps are initially ambiguous, but are eventually resolved by the network often well before the hand makes contact with the object. Figures 3 and 4 show examples of power and precision grasps generated by the simulator (see section 2.1 and [Oztop & Arbib, 2002] for grasp simulator implementation) and the time course of the network's output unit activity for each grasp. In these simulations there was no auditory component to the grasps. The results of these simulations show that the recurrent neural network functions just as effectively in grasp recognition as the feed-forward implementation developed in [Oztop & Arbib, 2002] and predicts a similar time course and pattern of mirror neuron activation for ambiguous grasps.

3.2 Audio-Visual Mirror Neurons

We tested the performance of the network in classifying an observed grasp under audio only, visual only, and audio-visual input conditions. Under each condition the simulated grasp and object were the same, but the arm and object's visibility and the action's audibility varied. In the audio-visual input condition, the audio input served to slightly strengthen the output activity. In the audio input only condition, the network correctly identified the grasp type associated with the sound. In this condition the output unit activity was confined to the duration of the audio input activity (see Figure 5).

3.3 Hidden Grasp Simulations

To simulate a hidden grasp, the object was visible to the network for the first 5 time steps of the grasps and was then set to invisible. The hand was visible for the first 26 time steps, and was then set to invisible as it reached to grasp the object behind the screen. During these simulations, no auditory information was presented to the network. The initial presentation of the object allowed its position and affordance information to be stored in working memory. The object and hand working memory traces utilizing dynamic remapping to update the wrist position were sufficient for the network to correctly recognize a hidden grasp (see Figure 6). Pantomimed grasps were simulated by setting the hand visible and object invisible for the whole grasp. To simulate a hidden pantomimed grasp, the hand was visible during the same time periods as in the hidden grasp, but the object was set invisible for the whole grasp. Neither of these methods allowed a trace of the object location and affordance information to be stored in working memory, and therefore the network correctly did not respond to either pantomimed grasp condition.

4 Discussion

We have shown that the monkey grasp-related mirror system can be modeled as a recurrent artificial neural network trained on self-generated grasps. The addition of audio input, working memory, and dynamic remapping give the network more flexibility in action recognition, allowing it to correctly recognize actions given only their sound and when the final component of the action is hidden. The ability to recognize invisible actions may allow primates to effectively monitor the

actions and infer the intentions of their peers in crowded environments. This capacity for action recognition is important for planning future actions and may underlie primate action selection processes in social situations.

4.1 Audio-Visual Mirror Neurons

We have shown that a recurrent neural network associating observed hand-object relation trajectories with motor programs can incorporate signals from other modalities by Hebbian association. This allows the monkey mirror neuron system to function as a multi-modal, actor-invariant representation of action, rather than a simple associator of visual and motor signals. If has been argued that Broca's area is the human homologue of area F5 in the macaque [Rizzolatti & Arbib, 1998] and that human language arose from a gesture based system which was later augmented with vocalization [Arbib, 2005]. These multi-modal mirror neurons may have allowed arbitrary vocalizations to become associated with communicative gestures, facilitating the emergence of a speech-based language from a system of manual gestures. If this is indeed the case, the development of audio-visual mirror neurons may have implications for the recognition of communicative actions and ground the multi-modality of language.

4.2 Inferring Hidden Actions

The results of these simulations show that the addition of a working memory with dynamic remapping of its representations is sufficient to infer the result of an action whose final component is hidden. The monkey mirror system is capable of inferring the final result of grasp given the initial sight of an object and a preshaped hand directed towards it even when the object and hand are subsequently obscured. It is not clear whether or not this system infers the actual outcome of the action, or the actor's intent in executing it. It has been proposed that mirror neurons are a part of, or a precursor to a simulation theory of mind-reading [Gallese & Goldman, 1998]. [Kuroshima *et al*., 2002] showed that capuchin monkeys can learn to infer whether or not a human knows the location of an object. We propose an experiment to test the involvement of the monkey mirror neuron system in inferring mental states and beliefs. Experimenter A would show an object to the monkey and then conceal it inside a box. The monkey would then observe experimenter B remove the object from the box after experimenter A leaves the room. At this point the monkey should know that the object is not inside the box, but that experimenter A believes that it is. Now if experimenter A returns and reaches inside the box, the monkeys grasp-related mirror neurons should discharge if they recognize intention, because experimenter A believes the object is inside the screen and intends to grasp it. If however, the mirror neurons code the inferred result of the executed action, they should be silent because the monkey knows that the object is not in the box. The results of this experiment could yield insight into the possible involvement of mirror neurons in primate theory-of-mind.

References

[Arbib, 2005] Arbib, M.A. From Monkey-like Action Recognition to Human Language: An Evolutionary Framework for Neurolinguistics. *Behavioral and Brain Sciences*, (In press).

[Arikuni *et al*., 1988] Arikuni, T., Watanabe, K., Kubota, K. Connections of area 8 with area 6 in the brain of the macaque monkey. *J. Comp. Neurol*., 277(1): 21-40, 1988.

[Deacon, 1992] Deacon, T.W. Cortical connections of the inferior arcuate sulcus cortex in the macaque brain. *Brain Research*, 573(1): 8-26, 1992.

[Dominey & Arbib, 1992] Dominey, P.F., Arbib, M.A. A cortico-subcortical model for generation of spatially accurate sequential saccades. *Cerebral Cortex*, 2: 153-175, 1992.

[Fagg & Arbib, 1998] Fagg, A.H., Arbib, M.A. Modeling Parietal-Premotor Interactions in Primate Control of Grasping. *Neural Networks*, 11: 1277-1303, 1998.

[Gallese *et al*., 1996] Gallese, V., Fadiga, L., Fogassi, L., Rizzolatti, G. Action recognition in the premotor cortex. *Brain*, 119: 592-609, 1996.

[Gallese & Goldman, 1998] Gallese, V., Goldman, A. Mirror neurons and the simulation theory of mind-reading. *Trends Cognit. Sci*., 2: 493501, 1998.

[Giles *et al*., 2001] Giles, C., Lawrence, S., Tsoi, A.C. Noisy Time Series Prediction using Recurrent Neural Networks and Grammatical Inference. *Machine Learning*, 44: 161-184, 2001.

[Hertz *et al*., 1991] Hertz, J., Krogh, A., Palmer, R.G. *Introduction to the Theory of Neural Computation*. Addison Wesley, 1991.

[Jones, 1992] Jones, M.J. *Using Recurrent Networks for Dimensionality Reduction* MIT, 1992.

[Kuroshima *et al*., 2002] Kuroshima, H., Fujita, K., Fuyuki, A., Masuda, T. Understanding of the relationship between seeing and knowing by tufted capuchin monkeys (Cebus apella) *Animal Cognition*, 5(1): 41-48, 2002.

[Koller *et al*., 2002] Koller, E., Keysers, C., Umilta, M.A., Fogassi, L., Gallese, V., Rizzolatti, G. Hearing Sounds, Understanding Actions: Action Representation in Mirror Neurons. *Science*, 297(5582): 846-848, 2002.

[Morel *et al*., 1993] Morel, A., Garraghty, P.E., Kaas, J.H. Tonotopic organization, architectonic fields, and connections of auditory cortex in macaque monkeys. *J. Comp. Neurol*., 335(3): 437-459, 1993.

[Oztop & Arbib, 2002] Oztop, E., Arbib, M.A. Schema design and implementation of the grasp-related mirror neuron system. *Biological Cybernetics*, 87(2): 116-140, 2002.

[Parker, 1995] Parker, L.E. The effect of action recognition and robot awareness in cooperative robotic teams. Proceedings of the 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 95), 1995.

[Rauschecker *et al*., 1995] Rauschecker, J.P., Tian, B., Hauser, M. Processing of complex sounds in the macaque nonprimary auditory cortex. *Science*, 268(5207): 111-114, 1995.

[Rizzolatti *et al*., 1996] Rizzolatti, G., Fadiga, L., Gallese, V., Fogassi, L. Premotor cortex and the recognition of motor actions. *Cogn. Brain Res*., 3: 131-141, 1996.

[Rizzolatti & Arbib, 1998] Rizzolatti, G., Arbib, M.A. Language Within Our Grasp. *Trends in Neuroscience*, 21: 188-194, 1998.

[Romanski *et al*., 1998] Romanski, L.M., Bates, J.F., Goldman-Rakic, P.S. Auditory belt and parabelt projections to the prefrontal cortex in the Rhesus monkey. *J. Comp. Neurol*., 403(2): 141-157, 1998.

[Schaal, 1999] Schaal, S. Is imitation learning the route to humanoid robots? *Trends Cognit. Sci*., 3: 233-242, 1999.

[Steijvers & Grunwald, 1996] Steijvers, M., Grunwald, P. A Recurrent Network that performs a Context-Sensitive Prediction Task. In *Proceedings from the Eighteenth Annual Conference of the Cognitive Science Society*, 335339, 1996.

[Umilta *et al*., 2001] Umilta, M.A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., Rizzolatti, G. I know what you are doing: a neurophysiological study. *Neuron*, 31(1): 155-165, 2001.

[Werbos, 1990] Werbos, P.J. Backpropagation through time: what it does and how to do it. In *Proceedings of the IEEE*, 78(10): 1550-1560, 1990.

Estimation of Eye-pupil Size during Blink by Support Vector Regression *

Minoru Nakayama
Tokyo Institute of Technology
CRADLE (The Center for Research and Development of Educational Technology)
Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
nakayama@cradle.titech.ac.jp

Abstract

Pupillography can be an index of mental activity and sleepiness, however blinks prevent its measurability as an artifact. A method of estimation of pupil size from pupillary changes during blinks was developed using a support vector regression technique. Pupil responses for changes in periodic brightness were prepared, and appropriate pupil sizes for blinks were given as a set of training data. The performance of the trained estimation models were compared and an optimized model was obtained. An examination of this revealed that its estimation performance was better than that of the estimation method using MLP. This development helps in the understanding of the behavior of pupillary change and blink action.

1 Introduction

Pupillography can be used as an index of mental activity and sleepiness [Kuhlmann and Böttcher, 1999; Beatty, 1982]. In particular, the eye sleepiness test, which is a kind of reading of the frequency power spectrum, can often be applied to measure the degree of tiredness in clinical observations or in industrial engineering situations. Mental activity and sleepiness are based on hi-level information processing, but pupil response only can not provide sufficient evidence of the process. Pupillography can be used as a measurable index to understand human behaviour, however.

Most methods of measuring pupil size are based upon the image processing of the eye. Therefore, any 'eye-blink' problem can affect measurements due to the eye being obscured by the eye lid during 'blink periods'. Blinks are usually discussed as an artifact for temporal observations such as mean pupil sizes or results of frequency analysis [Nakayama and Shimizu, 2001; 2004]. To extract the change in mental activity, the temporal pupil size should be measured accurately without blink artifact.

To reduce the influence of blinks on pupillary change, a model for predicting pupil size has been developed. This model may aid understanding pupillary behaviour or implicit action. An estimation method was developed using a three layer perceptron as a kind of multi layer perceptron (MLP) [Nakayama and Shimizu, 2001; 2002]. The training data was prepared as a pair of temporal pupillary changes. One was the original measurement without blinks, and the other was modified by replacing some periods with artificial blinks. Here, artificial blinks are typical patterns of pupillary change during blinks. The MLP was trained by reproducing original pupil size without blinks from pupillary changes with artificial blinks [Nakayama and Shimizu, 2001; 2002].

This estimation method can be applied to various experimental pupil sizes, however accuracy is often an issue. One of the possible reasons might be the estimation process. The MLP with sigmoidal function was applied to the estimation according to the pupil response, based on the non-linear model [Takahashi *et al.*, 1976]. The activation function might not represent pupillary change sufficiently. An alternative network to the MLP is the radial basis function (RBF) network [Luo and Unbehauen, 1997; Bishop, 1995], and this provides a smooth interpolating function by using basis functions such as the Gaussian function. This estimation issue suggests a kind of regression. Currently the support vector regression (SVR) can be used as a more robust representation for the regularization or the extraction of feature space. It is also suggested that the Gaussian kernels tend to yield good performance [Smola and Schölkopf, 1998].

Another reason might be that the training data consists of artificial blinks. Due to the method of measuring pupil size, the correct size during a blink is never obtained because the pupil is covered by the eye lid. Therefore it is not easy to prepare appropriate training data for making estimations.

In this paper, a new estimation method, which consists of a SVR technique and an experimental pupillary change, was developed to improve estimation accuracy and to observe pupillary response.

The purposes of this paper are addressed as follows:

1. To prepare a training data set which consists of pupillary change with blink artifact for developing the estimation method.
2. To develop an estimation model by use of a support vector regression technique, and to evaluate the estimation

*This research was partially supported by the Ministry of Education, Sports, Culture, Science and Technology, Grant-in-Aid for Exploratory Research 16650208, 2004-2005.

performance in comparison with other methods.

2 Method

2.1 Periodic pupillary response

Estimation of pupil size during blink provides a possible pupil size from the temporal sizes. Therefore, training data as a prototype of pupil response consists of experimental pupil sizes in the blink and possible sizes. As already suggested, it is not easy to measure the possible size during blinks, so the size should be obtained by estimation. In this paper the periodic pupillary responses were observed to extract typical response patterns because the pupil accurately reacts to light stimulus as an eye pupil reflex, despite blink artifact and pupillary noise. It is easy to know overall change in response to the brightness change. The light reflex was applied to control pupillary change and to extract regularized responses. In a sense, observing the reflex reaction was not the main purpose of this experiment.

The bright stimuli consisted of four square waves ($T = 4, 3, 2, 1sec.$) and three triangular waves ($T = 2.2, 5.3, 10.7sec.$) in the range of 10 cd/m^2 to 80 cd/m^2. The duration of each stimulus was 40 seconds. This visual stimuli was presented on a 17 inch computer monitor. Three subjects (Subject no.1$\sim$3) who have normal visual acuity took part in this experiment. They were seated with their heads on a chin-rest which was positioned 50 cm from the monitor.

Figure 1 illustrates pupillary change in response to a bright stimulus from a square wave ($T = 4sec.$) of 3 seconds. The horizontal axis shows time and the vertical axis shows pupil size and brightness change. The bold line shows brightness, the series of "•" show experimental measures of pupil size. Pupil responses indicate light reflex reactions with time delays which are approximately 0.2$\sim$0.5 seconds [Utsunomiya, 1978]. The figure shows that pupil size decreases gradually with time delays after the change in brightness. There are two drops which are caused by the blink. The average blink rate in this experiment was 12.3 blinks per minute. Usually, a subject blinks about 20 times per minute [Tada *et al.*, 1991]. It seems that subjects have suppressed blinks during the experiment, however.

The pupillary responses to the stimuli were observed using an eye tracker (nac:EMR-8) with pupil size measuring capability. The pupil image is captured by a small camera placed between the display and the chin-rest. The camera does not prevent the subject from seeing the display because it is located lower than the viewing level. The captured pupil image is analyzed as an ellipse which has longer and shorter diameters. This analyzing equipment measures the longer diameter of an ellipse of the pupil at 60 Hz, and produces the raw data and the status code of the measurement. This equipment also measures the shorter diameter simultaneously, to monitor the aspect ratio of the longer and shorter diameters. If the eye lid covered a part of pupil and the value of the diameter was affected, the aspect ratio would be smaller than 1.0 because of the round shape of the pupil. The longer diameter is the horizontal length of the ellipse and the shorter diameter is the vertical length during blinks. Both diameter and aspect ratio decrease with the degree of coverage of the pupil by the eye lid. When the aspect ratio of the pupil is under $0.7 \sim 0.8$, the measuring error code is given as the output status [nac Corp., 1999]. This means that pupil sizes during a blink are detected by the equipment. The easiest way to estimate pupil size during blink is to replace the drop in pupil size with a pupil size which is the last valid measurement before the status registers an error code. This replacement algorithm and process are very simple. The estimation value can be replaced automatically according to the status code, and it is defined as the "Auto correction". This estimation is also illustrated in Fig. 1. There is no drop in pupil size, but the accuracy is still not sufficient, however, because the pupil size has already begun to decrease when the output status produces an error code.

The pupil size as a circle was calculated from the longer diameter of the ellipse. The pupil size is zero when the whole pupil is covered by the eye lid, such as during a blink. The pupil size is significantly different among individuals, then the size is standardized by individual average size. Pupil size was originally observed at 60 Hz, however the data was resampled at 30 Hz to compare the estimation performance of the previous method [Nakayama and Shimizu, 2001].

2.2 Training data

To obtain pupillary change without blink, the pupil sizes during blinks were interpolated manually for three participants by referring to the periodical pupillary change. Because the pupil's light reflex to changes in the brightness of the stimuli is mostly consistent, pupil size can be interpolated from regular responses in other cycles if the pupil size during blinks has dropped off during a cycle. The interpolated periods are longer than the area where the measuring status registers on error code. This corrected pupil size is also illustrated in Fig. 1 as the "Reference". This seems to be a more plausible method of measuring pupillary change during blink than the estimation of "Auto correction".

As a result, a pair of temporal pupillary changes with and without blinks was prepared. Figure 2 illustrates the estimation process which is a mapping. A target pupil size is generated from the pupillary change during the "drop–off period" of the blink. Two out of the three sets of participant data were assigned as training data and the remaining set of participant data was assigned as test data.

2.3 Pupil size estimation by use of SVR

The estimation function was derived from the training data by use of the support vector regression technique [Smola and Schölkopf, 1998]. The estimation procedure is similar to the one using MLP [Nakayama and Shimizu, 2001]. As displayed in Fig. 2, a sub-string of data $\mathbf{x}$ which consists of n components is taken stepwisely from the time sequence data. Here, k th $\mathbf{x}$, $\mathbf{x_k}$ is noted as follows:

$$\mathbf{x_k} = (x_{k-(n/2-1)}, \ldots, x_k, \ldots, x_{k+(n/2-1)})$$

The estimated pupil size $\hat{y_k}$ for the empirical size y_k at the time position k is reproduced from $\mathbf{x_k}$. This requires deriving the mapping from the experimental pupil size with blinks to a pupil size without blinks, which is termed the "Reference". Here, the mapping function is defined as f. Figure 2 notes

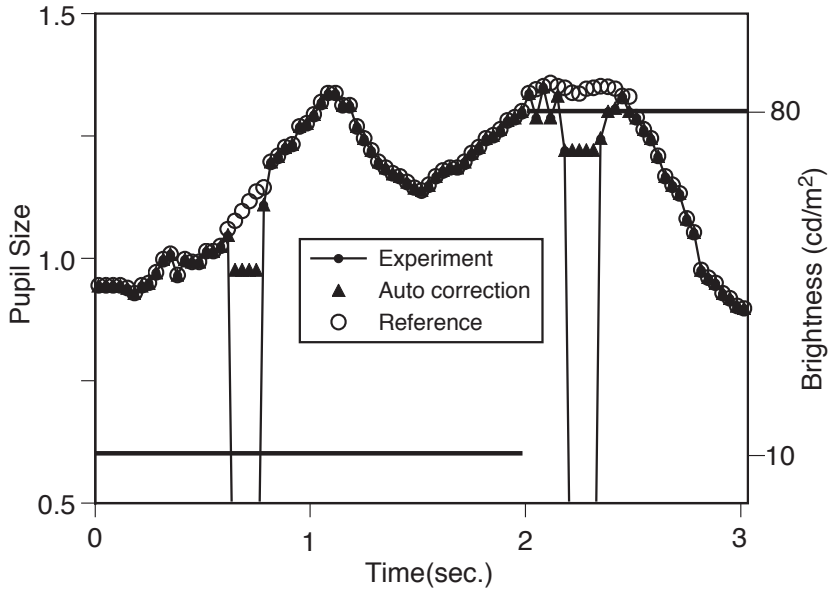


Figure 1: Light reflex pupillary change and training data (Bright stimulus $T = 4\ sec.$)

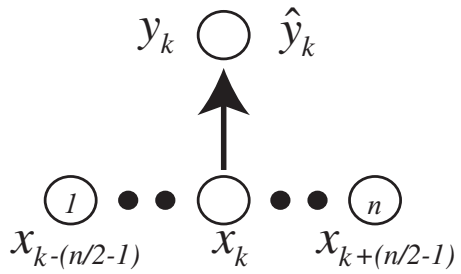


Figure 2: Relationship between experimental data ($\mathbf{x}$) and estimation target (y)

the mapping function from $\mathbf{x}$ to $f(\mathbf{x})$. The required mapping function f can provide an interpolated pupil size from $\mathbf{x}$ where $\mathbf{x}$ includes the zero value component x_i as blinks.

The number of the training data sets (x_i, y_i) is l, and the mapping function f is defined by linear regression as follows [Collbert and Bengio, 1998]:

$$f(\mathbf{x}) = (\mathbf{w} \cdot \mathbf{x}) + b$$

To estimate the function f with a precision of ϵ, the minimization problem is defined as introducing the geometric margin $\frac{1}{\|\mathbf{w}\|^2}$, as follows:

$$\frac{1}{2}\| \mathbf{w} \|^2 + C\sum_{i=1}^{l} |y_i - f(\mathbf{x}_i)|_\epsilon$$

where $\| \cdot \|^2$ is the Euclidean norm, $\frac{1}{2}\| \mathbf{w} \|^2$ is a regularization factor, C is a fixed constant, and $| \cdot |_\epsilon$ is the ϵ-insensitive loss-function.

$$|z|_\epsilon = max\{0, |z| - \epsilon\}$$

Here, one can introduce slack variables ξ, ξ^* to support vector "soft-margin" loss function, and then this is written as the minimization problem of τ [Collbert and Bengio, 1998].

$$\tau(\mathbf{w}, \xi, \xi^*) = \frac{1}{2}\| \mathbf{w} \|^2 + C\sum_{i=1}^{l}(\xi + \xi^*)$$
$$((\mathbf{w} \cdot \mathbf{x}_i) + b) - y_i \leq \epsilon + \xi_i$$
$$y_i - ((\mathbf{w} \cdot \mathbf{x}_i) + b) \leq \epsilon + {\xi_i}^*$$
$$\xi, {\xi_i}^* \geq 0.$$

To generalize as non-linear regression, a kernel $k(\cdot)$ is denoted non-linear transform $\Phi(x)$ for the feature vector $\mathbf{x}$. This procedure is the so-called *kernel trick*. Introducing Lagrange multipliers $\alpha_i, (i = 1, \ldots, l)$, this is the minimization problem of the objective function, as follows:

$$\frac{1}{2}\sum_{i=1}^{l}(\alpha_i - {\alpha_i}^*)(\alpha_j - {\alpha_j}^*)k(\mathbf{x}_i, \mathbf{x}_j)$$
$$- \sum_{i=1}^{l}(\alpha_i - {\alpha_i}^*) + \epsilon\sum_{i=1}^{l}(\alpha_i + {\alpha_i}^*)$$
$$\sum_{i=1}^{l}(\alpha_i - {\alpha_i}^*) = 0$$
$$0 \leq \alpha_i, {\alpha_i}^* \leq C$$

Then, function f is written as follows:

$$\mathbf{w} = \sum_{i=1}^{l}(\alpha_i - {\alpha_i}^*)\Phi(\mathbf{x}_i)$$

$$f(\mathbf{x}) = \sum_{i=1}^{l}(\alpha_i - {\alpha_i}^*)k(\mathbf{x}_i, \mathbf{x}) + b$$

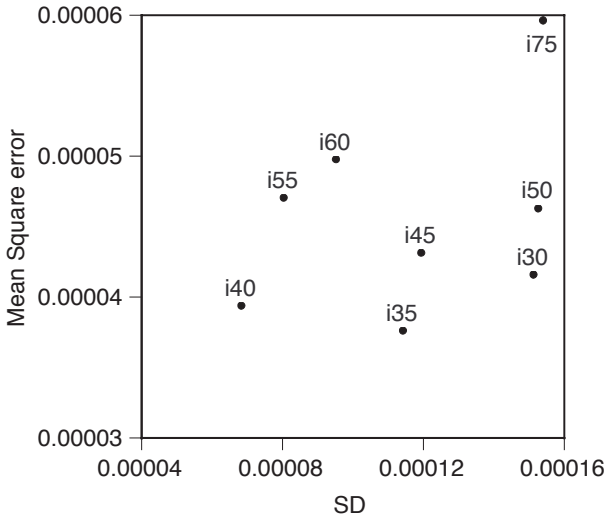


Figure 3: Mean and SD of square errors

In this paper, the Gaussian kernel in the introduction is introduced as follows:

$$k(\mathbf{x}, \mathbf{x}') = exp(- \parallel \mathbf{x} - \mathbf{x}' \parallel^2 /2\sigma^2)$$

Example l is given by the amount of training data. To obtain the optimized model, a dimension n of $\mathbf{x}$, and a precision $\epsilon(eps)$ and $\sigma(STD)$ of Gaussian kernel should be derived.

The practical calculation was conducted using the *SVMTorch* package [Collbert, 2000], and parameters were optimized.

3 Result

3.1 Reproducing performance

To derive the optimal condition, the performance which was reproduced, was examined according to the parameters.

The dimension n in Fig. 2 was examined across eight conditions: $n = 30, 35, 40, 45, 50, 55, 60, 75$ when the precision was set to $\epsilon(eps) = 0.01$. While in this condition, $\sigma(STD)$ of the Gaussian kernel function was controlled from 0.4 to 8.0 by 0.4 increment steps. For example, the amount of training data was $l = 15,416$ in the condition of $n = 35$.

After the training of the support vector regression with *SVMTorch II*, the performance for reproducing the training data was examined. The mean square error (MSE) and the standard deviation (SD) of errors were compared across the training conditions. The least mean square error and the standard deviation of error for each dimensional condition was summarized in Fig. 3 and labeled as input dimensions (i-n). The vertical axis shows the mean square error, the horizontal axis shows the SD of errors. As shown in Fig. 3, the least mean square error was $n = 35$. Where $n = 35$, the least MSE was $\sigma(STD) = 2.4$, and the number of *Support Vectors* was 1,380.

3.2 Estimation result

The trained model was applied to experimental data for Subject 3 in Fig. 1. The output of the model was illustrated as SVR in Fig. 4. According to the temporal pupillary change which was reproduced by SVR, all pupil sizes during blinks were replaced with plausible assumptions.

Another estimation using the MLP model [Nakayama and Shimizu, 2001], which was developed previously with empirical pupil sizes and artificial change for the blink, was also illustrated in the same format in Fig. 4. Some estimation sizes during blink were higher than the plausible sizes. Comparing the two estimations in this figure, the estimation with SVR seems more appropriate.

3.3 Estimation performance

To evaluate estimation performance of pupil size during blink, MSEs for the test data set were compared across the following four estimation methods. Here, the error was defined as the difference between the estimation size and the "Reference" which was determined, as was the training data.

1. Experimental measured size including blink influence (Exp)
2. Keeping previous valid size during blink (Auto)
3. Estimation size by MLP which was trained with artificial blinks (MLP)
4. Estimation size by SVR which was trained with the above training data (SVR)

For the estimation using SVR, the performance was compared across precision $\epsilon(eps) = 0.001, 0.005, 0.01, 0.05$. The parameter $\sigma(STD)$ was given for each precision condition according to the least square error for the training data. In general, the square error of the reproduction decreases as the precision $\epsilon(eps)$ becomes smaller.

Total square error and square error during blink periods for the test data set were summed up in each condition. Those errors were summarized in Fig. 5. The vertical axis indicates the total sum of square error, and the horizontal axis indicates the sum of the square error during blinks. Both axes are shown in logarithmic scale. For the experimental data including blinks, the total sum of the square errors resulted in drops in blinks. Therefore both errors coincided and were the largest value. Estimation performance of MLP was comparable to "Auto" condition.

When the performance of SVR was compared across precision parameters, the total square error in the condition of $\epsilon(eps) = 0.01$ was the least. According to the test result, parameters of the optimized condition are the input dimension $n = 35$, a parameter of the Gaussian kernel $\sigma(STD) = 2.4$, and a precision $\epsilon(eps) = 0.01$.

As a result, the total sum of the square error decreased to 25% of MLP, and 47% of "Auto". Also, the sum of the square error during blinks decreased to 43% of MLP, and 28% of "Auto". It is interesting that the sum of square error during blinks decreases with larger $\epsilon(eps)$.

3.4 Application to another data set

To examine the pupil size estimation possibility of the trained model, other experimental data as well as the test data was used. Pupillary change was measured in an experiment which gave oral calculation tasks to a subject while visual stimulus

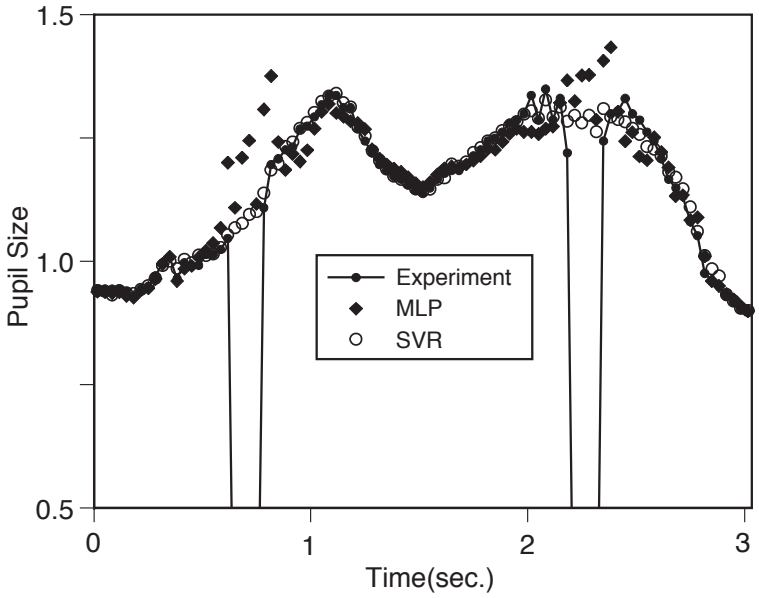


Figure 4: Estimation results using SVR and MLP

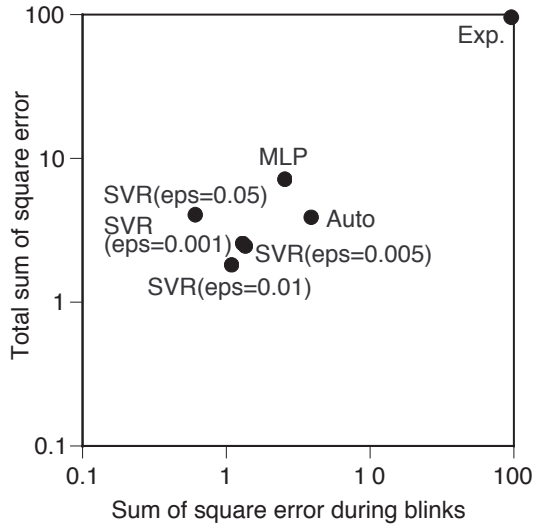


Figure 5: Square error change with $\epsilon(eps)$

such as the subsequent ocular task was displayed [Nakayama *et al.*, 2002]. Experimental pupil size which was measured at 30 Hz was illustrated for 10 seconds in Fig. 6 as Exp. The horizontal axis shows time and the vertical axis shows pupil size, and also drops show blinks.

The estimation result using the above trained model was overlapped in Fig. 6, as SVR. The SVR indicates the same pupil size without blink periods and gives possible sizes during blinks. However, blinks affect estimation sizes before or after blink periods. As the blink often widely influences pupil size before or after the blink, it is not easy to select the target period for estimation. Another reason is the difficulty in discriminating between correct and incorrect pupil sizes. Some irregular pupil sizes are displayed in Fig. 6, but they have valid sizes in the correct range. These will be the subjects of further study.

4 Summary

The estimation method of pupil size during blinks was developed using a support vector regression technique, while the training data was prepared from pupillary responses for 7 periodical brightness changes. According to the periodical pupillary changes, pupil sizes during blink were given manually, to prepare a pair of possible pupil sizes and empirical data. The parameters for the support vector regression technique were optimized in the training and test processes. The estimation performance was the highest amongst the proposed methods. This model can be applied to other pupillary observations which were conducted as part of an experiment using different subjects for a different purpose.

The model could simulate human eye pupil and blink, therefore there is a possibility to obtain the behavior of pupillary change and blink action. In particular, it may be possible to extract some features of pupil action as support vectors. Therefore, analysis of the support vectors and the relationship between pupil action and these support vectors should be conducted. The examination of these points will be the subject of further study.

References

[Beatty, 1982] Jackson Beatty. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2):276–292, 1982.

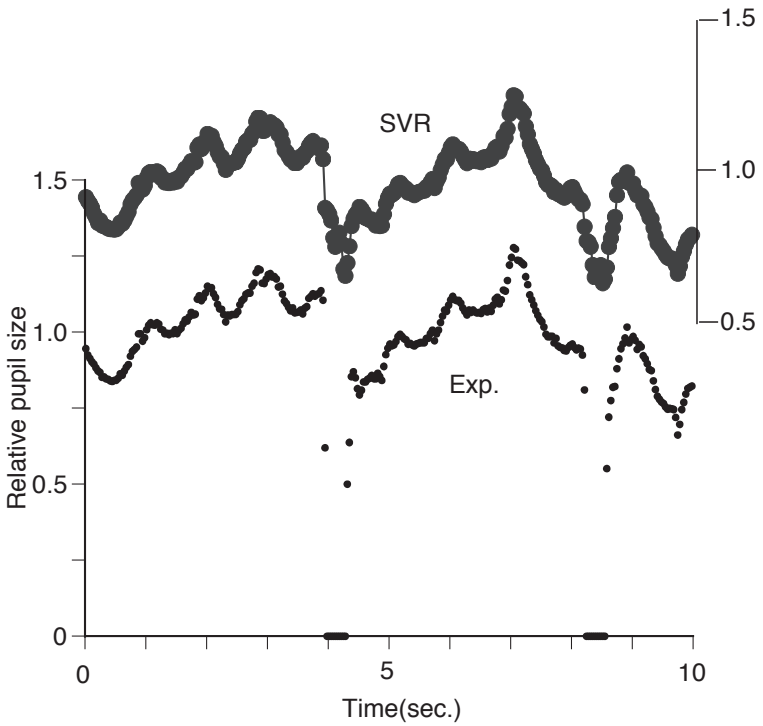


Figure 6: An application result

[Bishop, 1995] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK, 1995.

[Collbert and Bengio, 1998] Roran Collbert and Samy Bengio. Svmtorch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, 1:143–160, 1998.

[Collbert, 2000] Roran Collbert. *SVMTorch II package*. http://www.idiap.ch/learning/SVMTorch.html., 2000.

[Kuhlmann and Böttcher, 1999] Jochen Kuhlmann and M. Böttcher, editors. *Pupillography: Principles, Methods and Applications*. W. Zuckschwerdt Verlag, München, Germany, 1999.

[Luo and Unbehauen, 1997] Fa-Long Luo and Rolf Unbehauen. *Applied Neural Networks for Signal Processing*. Cambridge University Press, New York, USA, 1997.

[nac Corp., 1999] nac Corp. *EMR-8 manual (in Japanese)*. Tokyo, Japan, 1999.

[Nakayama and Shimizu, 2001] Minoru Nakayama and Yasutaka Shimizu. An estimation model of pupil size for blink artifact in viewing tv program. *IEICE Trans.,*, J84-A(7):969–977, 2001.

[Nakayama and Shimizu, 2002] Minoru Nakayama and Yasutaka Shimizu. An estimation model of pupil size for 'blink artifact' and it's applications. In Michel Verleysen, editor, *Proceedings of 10th European Symposium on Artificial Neural Networks*, pages 251–256, Evere, Belgium, 2002. d-side.

[Nakayama and Shimizu, 2004] Minoru Nakayama and Yasutaka Shimizu. Frequency analysis of task evoked pupillary response and eye-movement. In Stephan N. Spencer, editor, *Eye Tracking Research and Applications Symposium 2004*, pages 71–76, New York, USA, 2004. ACM, ACM Press.

[Nakayama *et al.*, 2002] Minoru Nakayama, Koji Takahashi, and Yasutaka Shimizu. The act of task difficulty and eye-movement frequency for the 'oculo-motor indices'. In Stephan N. Spencer, editor, *Eye Tracking Research and Applications Symposium 2002*, pages 37–42, New York, USA, 2002. ACM, ACM Press.

[Smola and Schölkopf, 1998] Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. In *NeuroCOLT2 Technical Report Series, NC2-TR-1998-030*, http://www.nerurocolt.com, October 1998.

[Tada *et al.*, 1991] Hideoki Tada, Fumio Yamada, and Kyousuke Fukuda. *Psychological blink (in Japanese)*. Kita-Ouji-Shobo, Kyoto, Japan, 1991.

[Takahashi *et al.*, 1976] Kunitaro Takahashi, Nakaakira Tsukahara, Keisuke Toyama, Mitsuhiko Hisada, and Hiroshi Tamura. *Neural Networks and Biological Control (in Japanese)*. Asakura Shoten, Tokyo, Japan, 1976.

[Utsunomiya, 1978] Toshio Utsunomiya. *Biological Control Information System (in Japanese)*. Asakura Shoten, Tokyo, Japan, 1978.

Biorealistic Simulation of Baboon Foraging using Agent-Based Modelling

W. I. Sellers
University of Loughborough
Department of Human Sciences
Loughborough UK
W.I.Sellers@lboro.ac.uk

R. A. Hill
University of Durham
Department of Anthropology
Durham UK
r.a.hill@durham.ac.uk

B. Logan
University of Nottingham
School of Computer Science & IT
Nottingham UK
bsl@cs.nott.ac.uk

Abstract

We present an agent-based model of the key activities of a troop of chacma baboons (*Papio hamadryas ursinus*) based on data collected at the De Hoop Nature Reserve in South Africa. The construction of the model identified some key elements that were missing from the field data that would need be collected in subsequent fieldwork. The simulation results identified decisions concerning movement (group action selection) as having the greatest influence on the outcomes. We analysed the predictions of the model in terms of how well it was able to duplicate the observed activity patterns of the animals and the relationship between the parameters that control the agent's decision procedure and the model's predictions. The model predicts reasonable yearly average values for energy intake, time spent socialising and resting, and habitat utilisation, but is unable to account for month by month variation in the field data. However even at the current stage of model development we are able to show that, across a wide range of decision parameter values, the baboons are able to achieve energetic and social time requirements. This suggests that these particular animals may be influenced by other factors such as predation risk or thermal load in deciding their activity patterns.

1 Introduction

Where two activities cannot be performed simultaneously, animals are forced to schedule certain behaviours preferentially, such that costs may be incurred by the reduced opportunities to engage in other biologically important activities [Caraco, 1979]. As a consequence, determining how ecological and demographic constraints influence time budget allocation and scheduling decisions is a key issue underlying detailed understanding of primate socioecology. Agent-based modelling is a powerful tool for ecological modelling and is especially suitable for situations where individual strategy and planning may be important as is commonly assumed to be the case when considering primates. However for this technique to be useful it is necessary to confirm that the model is sufficiently complex to represent observed behaviour patterns and to identify the environmental and behavioural measures that need to be collected to (a) build a successful model and (b) validate the predictions of the model against experimental data. In this paper we consider these issues using a set of empirical data from a troop of chacma baboons (*Papio hamadryas ursinus*) with associated ecological data collected at the De Hoop Nature Reserve in South Africa. We present an agent-based model designed to simulate the key activities of the troop and analysed its predictions in terms of how well it was able to duplicate the observed activity patterns of the animals and also in terms of the relationship between the parameters that control the agent's decision procedure and the model's predictions.

The remainder of the paper is organised as follows. In section 2 we motivate our approach to agent-based modelling and our choice of group behaviour as the focus of this paper. In section 3 we briefly summarise the field data on which our model is based and in section 4 we outline our agent-based model and the decision procedure which the agents use to choose their activities. In section 5 we present the results of a Monte-Carlo sensitivity analysis of the parameters used in the agent's decision procedure. In section 6 we discuss the results and in section 7 we briefly outline some related work. In section 8 we conclude and outline directions for future work.

2 Agent-based modelling

Individual-based ecological models have been growing in importance over the last 20 years and it has been predicted that this reductionist approach will provide valuable insight into system wide properties [Lomnicki, 1992]. Early work in artificial life has shown that complex group behaviours such as flocking and following can be produced using simple rules applied to individuals [Reynolds, 1987]. Agent-based modelling is an extension of this approach where each individual retains information about its current and past states, and its behaviour is controlled by an internal decision process. An agent is a software system that perceives its environment and acts in that environment in pursuit of its goals. Agents integrate a range of (often relatively shallow) competences, e.g., goals and reactive behaviour, emotional state, memory and inference. In agent-based modelling, the agents are situated in a simulated environment, and are equipped with sensors with differing ranges and directional properties (e.g., smell, hearing, vision) and the ability to perform a range of actions

which change the state of the environment or the perceptible characteristics of the agent. The environment may contain passive objects (e.g., topography) and active objects and processes which change spontaneously during the course of the simulation (e.g., weather) and/or in response to the actions of the agents (e.g., food bearing plants).

The outcome of this process depends on the set of desires and goals within the individual, its current internal state, an internal world model, and sensory information. This reliance on individual choice makes this technique especially useful when dealing with intelligent organisms since it is likely that the optimal strategy for an individual depends on the strategies adopted by others in the group [Milinski and Parker, 1991]. The justification for this approach is that whilst the factors influencing the decisions made by an individual may vary as the environment changes, the decision process itself is likely to be conserved, and an agent with a robust decision procedure will demonstrate reasonable behaviour under a wide range of conditions. If we are confident that the decision procedure is robust, then we can use the behaviour of the agents to predict the behaviour of real populations and to explore the potential effects of situational changes: climate, food distribution and body size can all be altered and the effects on the agents' behaviour can be observed.

Agent-based modelling has become a popular technique for modelling social and spatial interactions in humans and non-human primates: virtual worlds populated by decision-making agents have been used to investigate topics as diverse as primate social hierarchies [Hemelrijk, 2002] and Mesolithic hunter-gatherer behaviour [Mithen, 1994]. However such computer simulations are not without their critics. John Maynard-Smith has famously described these approaches as "fact-free science" [Maynard-Smith, 1995]. To overcome such objections and to enable us to use this technique as a tool for exploring primate behavioural ecology, the models produced must be tested by using them to predict behaviours in a given population and comparing the predictions with field observations.

No model can accurately predict all aspects of primate behaviour. Even if it could, it is unlikely that it would be useful, as one of the primary functions of a model is abstract the key features of the system of interest. In this paper, we focus on the problem of action selection in groups, i.e., where an individual's action choice is constrained by the choices of other members of the group. Group living is a common strategy among mammals and is key to understanding the success of the primate order in general and early humans in particular. A great deal of ecological theory has been generated to investigate grouping strategies and to identify the optimal group size given various ecological parameters [Cheney, 1987; Dunbar, 1996]. We focus on measures such as range size, daily travel distance, energy and time budgets, as these are good candidates for testing agent-based approaches: they have measurable numerical values and so can be tested objectively, and they are highly dependent on the activities and choices of the individual within the population. Baboons (*Papio* spp.) are one of the most widely studied primate species and are ideal for studies of primate ecology since they often live in open, terrestrial habitats, and can be observed closely for long periods of time [Richard, 1985]. This means that there is a wealth of data available documenting most aspects of their behaviour in great detail. Many of these studies have managed to quantify the activity patterns of individuals both in terms of durations and also the costs and benefits of the activity. *Papio* spp. are found across most of sub-Saharan Africa [Jolly, 2001], at a range of altitudes, with attendant large changes in average rainfall and temperature. Thus they can be said to inhabit a wide variety of habitats and ecotypes, and studies have shown that their diet and foraging varies in response to environmental determinants [Hill and Dunbar, 2002].

Our long term aim is a robust model of baboon behaviour which is valid across a wide range of habitats and baboon species (including extinct species). Our methodology is to first build a model that can successfully predict the behaviour of a particular group of baboons and then attempt to generalise, conserving the decision procedure while tailoring the decision parameters to a particular species or habitat type. The work reported in this paper is the first step in this process, namely the modelling of a particular group of baboons in a particular habitat.

3 Field Data

The model is based on data from De Hoop Nature Reserve, a coastal reserve in Western Cape Province, South Africa. Vegetation is dominated by coastal fynbos, a unique and diverse vegetation type comprising Proteacae, Ericaceae, Restionaceae and geophyte species. Seven distinct habitat types were classified on the basis of vegetation structure within the home range of the baboons (Table 1: see [Hill, 1999] for detailed descriptions and further information on the ecology of the reserve).

De Hoop has a mean annual rainfall of 428 mm, with a mean annual temperature of $17.0°$C. Both rainfall and temperature show considerable seasonal variation and shade temperature in the summer months regularly exceeded $25°$C. Due to its southerly latitude, De Hoop also experiences significant day length variation (from 9.8 to 14.2 hours) that has important implications for the behavioural ecology of this population [Hill *et al.*, 2003].

The data presented here are for a 7-month period (June to December 1997) from a single troop of chacma baboons (*Papio hamadryas ursinus*) that ranged in size from 40 to 44 individuals over the course of the study. Data were collected by means of instantaneous scan samples [Altmann, 1974] at 30-minute intervals, with 2-4 adult males and 12 adult females sampled for a minimum of five full days each month. At each sample point, information was recorded on the identity, habitat type and activity state (feeding, moving, socialising or resting) of all visible individuals. Each scan lasted a maximum of 5 minutes. A more detailed description of the data collection methods is given in [Hill, 1999].

4 Agent-Based Model

The model consists of two components: the environment model and the baboon model. The environment model was

Habitat Type	Proportion of Range (%)	Bush Cover (%)	Tree Cover (%)	Food Availability	Shade Availability	Predation Risk
Acacia Woodland	15.8	55.8	34.4	High	Very high	High
Burnt Acacia Woodland	1.2	3.2	0.4	Low	Low	Intermediate
Burnt Fynbos	27.6	3.6	0.0	Low	Low	Intermediate
Climax Fynbos	25.7	54.0	3.4	Low	High	High
Grassland	11.0	1.6	1.2	Intermediate	Low	Low
Vlei	18.7	0.0	0.0	High	Very low	Low

Table 1: Home range composition, vegetation food availability, shade availability and predation risk of the major habitat types at De Hoop.

based on the 200×200m map grid used for field data recording, and consists of 660 cells within an area 5.4km by 8.4km. Each cell contains a mixture of the 6 primary habitat types found at De Hoop (Acacia Woodland, Burnt Acacia Woodland, Climax Fynbos, Burnt Fynbos, Grassland and Vlei) and may also include one or more 'special features': water sources, sleeping sites, and refuges (primarily cliffs). Each habitat type was characterised by a maximum food availability, energy intake rate when foraging and foraging on the move, and replenishment rate, and these varied month by month. For reasons of space, these values will be reported elsewhere. The actions of the agents affect the environment. For example, food consumed is depleted from the grid square containing the agent, and is replaced at the replenishment rate for the current simulation month for each of the habitat type(s) occurring in the grid square. The energy value of the food available was estimated at $13.98\,\mathrm{kJg}^{-1}$ [Stacey, 1986]. The environment model is illustrated in the simulator's graphical output shown in Figure 1.

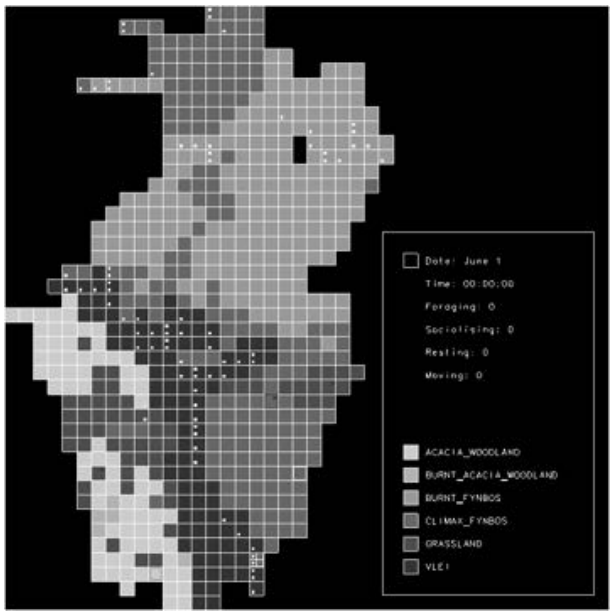


Figure 1: Graphical output from the simulator showing the habitat types and distributions.

The baboon model models each baboon as an agent with physical parameters based on well-known baboon physiology. In addition, each agent maintains an individual score for thirst, energy and social time. These scores function as 'drives' or 'desires' in biasing the agent's choice of preferred activity at each timestep.

At each timestep, each agent can choose one of four actions corresponding to the activities recorded for the baboons at De Hoop: feeding, moving, socialising or resting. In addition, an agent can perform an instantaneous drinking action which can be combined with any of the other activities (assuming the agent is in a cell which contains a water source). Each action has an associated energy cost. These were calculated using the formulae given in [Tucker, 1970] for an average adult female baboon with a body mass of 16.1 kg (the heavier males offset by the lighter infants and juveniles) and assuming that the baboons moved relatively slowly ($0.5\,\mathrm{ms}^{-1}$) since they customarily foraged whilst moving. Thus foraging uses 36.71 W; moving 50.59 W; socialising 64.04 W; and resting 34.63 W. In addition to its energy cost, each action updates the appropriate scores. Feeding causes food to be depleted from the grid square containing the agent and increases the agent's energy score depending on the type of food consumed. The agents also forage while moving, which depletes food from the grid square at a lower, travel-foraging, rate.[1] Socialising increases the agent's social time score by the length of the timestep. Drinking adds one to the agent's thirst score. Any action other than socialising causes the social score to decrease by the length of the timestep, and not drinking causes the agent's drinking score to decrease by the reciprocal of the timestep.

The agents have two hard constraints: they must return to a sleeping site to rest each night and they must drink (i.e., visit a grid square constraining a water source) at least once every 2 days. Otherwise they have 2 goals: to maintain their energy balance (i.e., to eat sufficient food to make up for the energy expended each day) and to spend 2 hours a day in social activity. Each agent is equipped with a simple decision procedure designed to allow it to exploit the habitat to achieve its goals. However the actions of each individual are constrained by actions selected by the other baboons in the manner explained below.

The model uses a fixed timestep of 5 minutes. At the end of each timestep each baboon chooses a preferred action to

[1] In some habitats there is no food to be gained by foraging on the move in certain months.

perform at the next timestep and whether it would prefer to move to allow it to perform the action more effectively. If the number of agents which vote to move is higher than a given threshold, V, then the whole group moves in the most commonly preferred direction. If the group does not move, agents which voted to move have the opportunity to choose an alternative action which can be performed in the current grid square. The agent then spends the next 5 minutes performing its chosen action and its scores in terms of energy balance, thirst and social time adjusted accordingly. Drinking is considered an instantaneous action that occurred whenever the agent occupied a grid square containing a water source.

The complete decision procedure can be summarised as follows:

```
Decision procedure:
    for each agent:
        if (!Automatic Action):
            Choose Preferred Action
            and Preferred Cell;
            if Preferred Cell != Current Cell:
                Vote to Move;

    if ((Votes to Move / no. agents) > V):
        for each agent:
            Move in the most commonly
            preferred direction;
    else:
        for each agent:
            Check Preferred Action;
            Perform Action;
```

`Automatic Actions` are the hard constraints (resting at night, drinking) which can preempt the choice of `Preferred Action` and `Preferred Cell`. For example, the requirement that the agents must return to a sleeping site to rest each night constrains the choice of `Preferred Cell` so that the agent can always reach a sleeping site in the time remaining before nightfall. If the `Automatic Action` step has not determined the agent's choice of action at this timestep, the agent's `Preferred Action` is determined using a weighted random function with weights proportional to the current desire to drink, forage and socialise. Desires are linear functions of the corresponding scores with gradients proportional to user defined relative importance values for each action: W_D (the relative importance of drinking), W_F (the relative importance of foraging), and W_S and the relative importance of socialising. These desire functions fall to zero when the target amount has been reached and when they are all zero the agent will opt to rest. By aiming to keep all scores at zero, the agents will drink on average once per day, and socialise on average 2 hours per day.

The agent will `Vote to Move` if it can perform its desired action more effectively in one of the neighbouring cells. This is determined by evaluating all the grid squares within the search radius, S, of the agent's current location. For drinking it is assumed that the agent knows where the nearest water source is, irrespective of search radius. The agent will vote to move if the best grid square is more than a user defined threshold better than the current square. In the case of foraging the threshold is denoted by T_F, and depends on the food availability, in the case of socialising and resting the thresholds (denoted by T_S, T_R) are a measure of predation risk. The votes for all the agents are counted and the group will only move if more than V vote in favour of moving. If fewer than V agents opt to move, then the agents which preferred to move choose their most preferred action for the current cell at the `Check Preferred Action` stage. This is because it is impossible to drink if there is no water in the current cell and undesirable to socialise or rest if the predation risk is greater than T_K. Finally all agents either move in chosen direction or get to `Perform Action`, which is whatever non-move action was decided upon after the preferred action was checked.

Parameter	Min	Max
V	0.1	0.9
S	200	2200
W_F	1	10
W_S	1	10
W_D	1	10
T_F	1	3
T_S	1	3
T_R	1	3
T_K	0	0.25

Table 2: Key parameters in the decision procedures showing the ranges used in the Monte-Carlo sensitivity analysis.

The difficulty is that we do not currently have suitable values for the parameters used in the decision process. Some we may be able to estimate empirically with more detailed field observations, but others are essentially unknowable. To overcome this we choose plausible ranges for each decision parameter and performed a Monte-Carlo sensitivity analysis [Campolongo *et al.*, 2000] where the simulation was repeated a large number of times and the values of the parameters randomly sampled from the range for each run. This allows us both to estimate the importance of a particular parameter on the outcome and to calculate the range of possible outcomes. The parameter ranges used in the analysis are shown in Table 2.

5 Results

The model was run 100,000 times sampling the decision parameters from Table 2 each time. Figure 2 (a) and (b) show the distribution of outcomes in terms of the goal states (daily energy intake and daily social time). These values are highly consistent between runs and almost always adequately high suggesting that almost no matter what combination of decision parameters we use the agents are able to achieve their goals. However if we look at the match between the model's predicted outcome and the recorded monthly activity summaries from the field data we can see that there is a very large amount of variation in the model's predictions and not particularly good agreement with the field data. This is shown in Figure 3 where the predictions in terms of time spent in different activities are compared with the experimental data.

Figure 4 which shows the time spent in different habitats shows a similar picture. In particular it shows that the baboons spent a great deal of time in the Burnt Fynbos in August (and to some extent in September). This is very hard

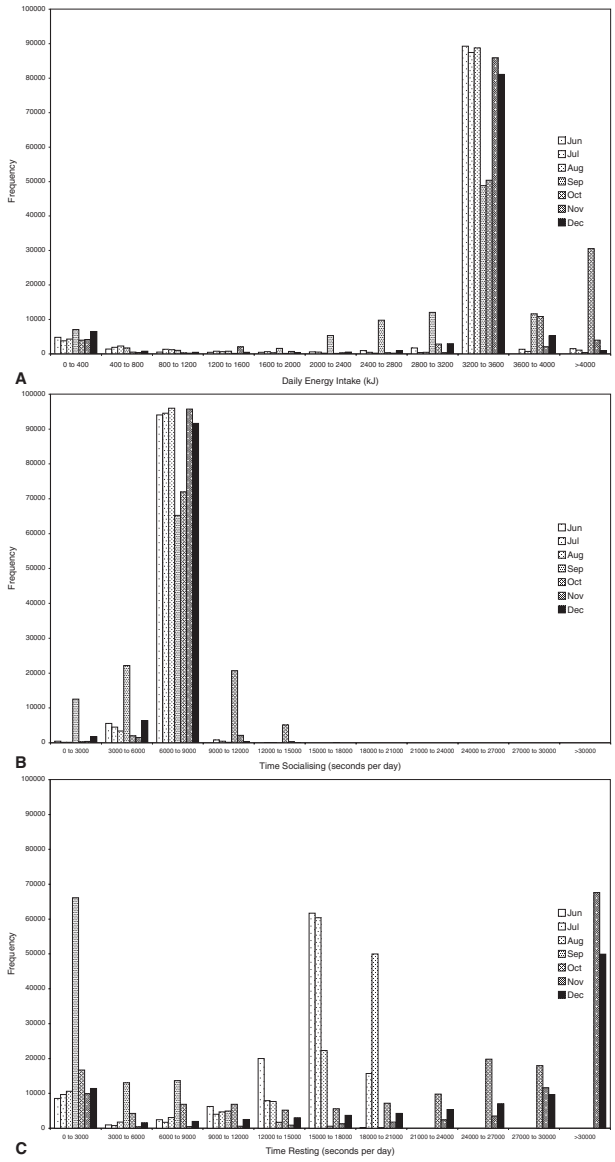


Figure 2: Frequency of the primary outcomes for 100,000 repeats of the simulation with the input parameters randomly sampled from the ranges in Table 2. (A) Daily energy intake (target approx. 3500 kJ depending on activity pattern); (B) Time spent socialising (target value is 7200 s); (C) Time spent resting.

to explain in terms of food availability and it is not surprising that it is not matched by the model. In terms of the whole year the model is much better able to match the observed findings.

Figure 5 shows the mean daily activity pattern for the whole year. Figure 6 shows the occupancy rates for different habitat types which also show a similar pattern, although the in the simulation the time spent in Acacia Woodland is consistently less than that spent by the actual baboons. The Burnt Fynbos usage over the whole year matches the baboon data even though it cannot mimic the August peak.

The linear effects of the decision parameters were analysed

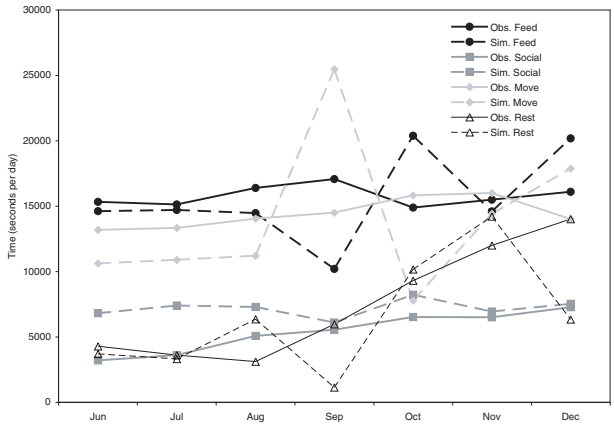


Figure 3: The daily duration of the 4 activities observed in the field and in the best matching run obtained in 100,000 repeats using randomly sampled decision parameters.

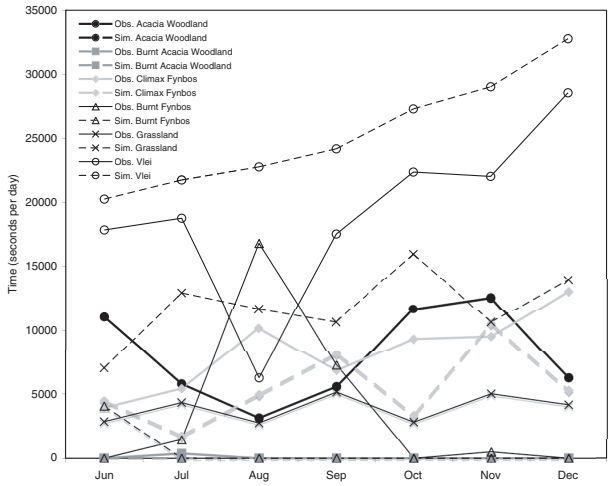


Figure 4: The daily amount of time each agent spends in a particular habitat observed in the field and in the best matching run obtained in 100,000 repeats using randomly sampled decision parameters.

using Stepwise Multiple Regression in SPSS. Table 3 shows the regression terms identified by this process. The clearest relationship is between the V parameter and the time spent moving. V also had a reasonably strong effect on the time spent foraging, the energy intake, and the time spent in Acacia Woodland. The time spent in Grassland is quite strongly influenced by the search radius, S. In no case does the second term add much to the overall relationship.

6 Discussion

The data show that the model is able to approximate the behaviour of the De Hoop baboon troop in general terms. However it is where the model and the real data differ that is most informative. It is commonly supposed that the requirement to obtain sufficient food is the key factor that produces primate movement. The energetic aspects of the model are probably

Dependent Variable	Predictor 1	R square 1	Predictor 2	R square 2
Time Foraging	V	0.412	W_F	0.423
Time Resting	T_F	0.043	W_F	0.060
Time Moving	V	0.755	T_F	0.827
Time Socialising	W_F	0.018	V	0.028
Energy Intake	V	0.255	W_F	0.264
Acacia Woodland	V	0.368	T_F	0.490
Burnt Acacia Woodland	V	0.111	S	0.187
Burnt Fynbos	V	0.061	W_F	0.063
Climax Fynbos	V	0.063	W_D	0.069
Grassland	S	0.304	V	0.330
Vlei	V	0.054	S	0.064

Table 3: The first two terms of a linear stepwise regression using the Monte Carlo parameters as independent variables and the behavioural outcomes as the dependent variables using the combined 100,000 runs as the data source.

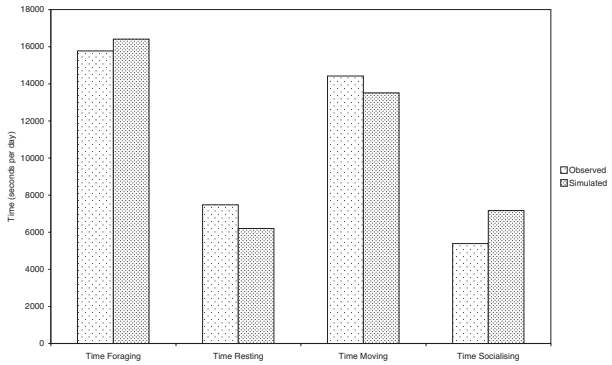


Figure 5: The daily duration of the 4 activities averaged over the seven month sample time observed in the field and in the best matching run obtained in 100,000 repeats using randomly sampled decision parameters.

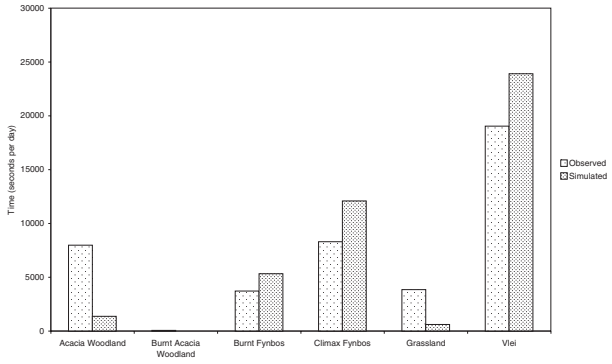


Figure 6: The daily amount of time each agent spends in a particular habitat averaged over the seven month sample time observed in the field and in the best matching run obtained in 100,000 repeats using randomly sampled decision parameters.

its most reliable features: we have reasonably good data for daily food requirements and nutrition physiology. The fact that even this simple model is able to obtain sufficient food (there is no planning of optimal routes or intelligent choice of when to feed) suggests that it is actually relatively easy for these animals to obtain sufficient food and that there must be other drives that have a strong influence on movement. Hill [1999] showed that predation risk was an important influence on the habitat choice of baboons and this is one area where the current model is weak. The model contains no free-ranging predators and predation risk is considered only in relation to resting and grooming. Risk is taken to be a function of visibility rather than a more complex model of predator-prey interactions. Other factors that might be important include the fact that, although energy is plentiful in the environment, particular nutritional components such as protein may be much rarer. This might explain the high usage of the otherwise undesirable Burnt Fynbos habitat in August and September, if the food items available there (mostly subterranean tubers and roots) contain specific, rare dietary elements. More detailed experimental and observational data will be needed to answer this, and the dietary component of the model will need to become more complex accordingly.

The Monte-Carlo analysis revealed that the model has very little clear, linear dependence on any of the input parameters. The V parameter is involved only in the decisions to move and its influence it almost certainly entirely due to its strong effect on the time spent moving. This is one of the least realistic parts of the model since clearly baboons do not actually vote to move as the simulated baboons do.[2] However they do always move in more or less coherent groups (it is extremely hazardous to be a lone baboon) so some sort of coordination mechanism must be at work and this is one area where the model would benefit from elaboration. Large values of S re-

[2]Conradt and Roper [2003] have shown that under certain assumptions, "democratic" decision-making results in lower costs to the group as a whole than "despotic" decision-making. They give as empirical examples of 'voting' behaviours the use of specific body postures, ritualised movements, and specific vocalisations, whereas 'counting votes' includes adding-up to a majority of cast votes, integration of voting signals until an intensity threshold is reached, and averaging over all votes. They cite anecdotal reports of voting behaviours in baboons where a simple majority determines changes in group activity based on movement, or a majority of adults or adult males decide on the direction of travel based on body orientation or position on a resting rock.

flect a better knowledge of the environment. Low values of S lead to more time being spent in grassland, suggesting that the agents are then failing to find better habitat in these cases.

7 Related Work

Agent and individual-based modelling is an increasingly popular approach to the study of primates. In this section we briefly review some of this work and sketch its relationship to the model described in this paper.

Robbins and Robbins [2004] have developed a model to simulate the growth rate, age structure and social system of mountain gorillas in the Virunga Volcanoes region. The model uses a one year time step and is based on the probabilities of life history events (birth rates, mortality rates, dispersal patterns etc.) as determined by census data from habituated research groups of gorillas. Hemelrijk [2002] presents a model of primate social behaviour in which agents have two tendencies: to group and to perform dominance interactions. By varying group cohesion she shows that denser grouping can induce female dominance over males. Bryson and Flack [2002] have used an agent-based model to investigate primate social interactions. The agents are represented as 2D rectangles in a walled enclosure which alternate between two behaviours: grooming neighbours and wandering (feeding in relative isolation). They investigated the effect of a 'tolerance behaviour' on the amount of time spent grooming.

For want of better terminology, we can distinguish between individual-based and agent-based models. An *individual-based* model takes individuals as the basic unit and tracks them without the individuals interacting in a meaningful way. We reserve the term *agent-based* for those individual-based models in which the individuals interact with an environment and/or each other.[3] For example, the model proposed by Robbins and Robbins is individual-based in that the gorillas don't interact with an environment model (or each other) and the only decisions the gorillas make as individuals is whether to move to a new group. In contrast the models by Hemelrijk and Bryson and Flack are agent-based in that they focus on the interaction of the individuals in the simulation. In particular, both models explicitly take into account the spatial position and orientation of individuals: in the Hemelrijk model, cohesiveness is determined by the 'SearchAngle', the angle by which an agent will rotate to locate other agents when there are none in sight; in the Bryson and Flack model, grooming requires being adjacent to and properly aligned with an agent.

Our approach is intermediate between individual-based and agent-based: baboons are modelled as individuals which choose actions and interact with their environment based on their individual state, but their interactions with each other are limited to group level decisions, specifically whether to move, and the constraints this places on their individual choice of action. To the best of our knowledge this integration of individual and group level action selection (where all the members of the group participate in the selection and execution of a common action) has not been addressed in previous work.

[3]Note that this usage is not consistent with that in the literature generally or even the papers summarised in this section.

There is also a substantial body of work on joint action in AI, for example [Grosz and Sidner, 1990; Cohen and Levesque, 1991; Tambe, 1997]. However this work has tended to view actions by individuals within a group as directed towards the achievement of a joint intention, with each agent committing to performing a (possibly different) action from a shared or team plan, rather than the selection of an action which is performed by all agents but which only serves the interests of a subset. It seems unlikely that baboons have joint intentions, or the shared plans and models of teamwork necessary to achieve them.

8 Conclusion

This study shows the potential value of agent-based modelling in primatology. It clearly demonstrates that, for this population, factors other than food are important for ranging. The construction of the model has identified key areas where the available field data is missing, and so is extremely useful for planning future studies. It also shows the non-linear nature of the problem and indicates useful ways that the model could be elaborated to investigate more complex issues, such as predation and planning.

The fact that the model is able to match the yearly activity and occupancy profiles suggest that even a simple model is perfectly adequate to simulate primate behaviour recorded at this time scale. However it is clearly unable to match the detailed activity at even a monthly, let alone daily or hourly time scale, although the results presented here do suggest that this should be possible. Some of the disparity between the model's predictions and the field data may be attributable to the fact that the field data actually represent a subset (scan samples at 30 minute intervals on 5 days per month) of the complete monthly behavioural profile of the baboons. In contrast, the model simulates the behaviour of the baboons every five minutes on every day each month. As a consequence, the field data are more susceptible to stochastic sampling variation where 'atypical' behaviour patterns could produce misleading monthly averages. The fact that the model matches the long-term yearly averages where such effects are minimised, therefore, is extremely encouraging.

There are a number of areas where additional detail could be beneficially added to the model. Firstly, the incorporation of a full diet model may be essential. This would be easy in modelling terms but difficult in terms of validation, since it would require much more detailed chemical and calorific analysis of what the baboons actually eat in different areas. Secondly, since it seems likely that predation is a major driving force of primate ranging behaviour, this would need to be incorporated specifically in the model. Fortunately this is precisely where agent-based modelling reveals its power and generality, since the predators can be modelled as agents themselves. The difficulty here is that we know considerably less about the behaviours of any of the predator species than we do the prey animals, so that validation may be extremely difficult. Thirdly, it seems likely that primates, and in particular baboons because of their larger than normal brains (for equivalent sized mammals) [Jerison, 1973], do have some sort of a mental map of their home range and do plan their

daily activities to some extent. It is obviously almost impossible to know how a baboon might view the world but an agent-based model is an ideal way of investigating possible approaches and can certainly quantify the costs and benefits associated with various levels of planning.

It would also be interesting to extend the model to explore the relationship between individual and group level action selection in more detail. For example, it would be straightforward to incorporate a weighted voting scheme in which the votes of some individuals have a greater effect on action choice (and in the limit some subset of individuals determines group actions). However it would be more interesting to try to model the emergence of group level action selection from the sum of interactions between individual agent's action choices (i.e., without an explicit voting scheme). This would require a much finer grained model of baboon sensing and behaviour, and a greater time resolution of the model.

References

[Altmann, 1974] J. Altmann. Observational study of behaviour: sampling methods. *Behaviour*, 49:227–267, 1974.

[Bryson and Flack, 2002] Joanna J. Bryson and Jessica C. Flack. Action selection for an artificial life model of social behavior in non-human primates. In Charlotte Hemelrijk, editor, *Proceedings of the International Workshop on Self-Organization and Evolution of Social Behaviour*, pages 42–45, Monte Verita, Switzerland, September 2002.

[Campolongo *et al.*, 2000] F. Campolongo, A. Saltelli, T. Sorensen, and S. Taratola. Hitchhikers' guide to sensitivity analysis. In A Saltelli, K Chan, and E M Scott, editors, *Sensitivity Analysis*, pages 15–47. Wiley, Chichester, 2000.

[Caraco, 1979] T. Caraco. Time budgeting and group size: a theory. *Ecology*, 60:611–617, 1979.

[Cheney, 1987] D. L. Cheney. Interaction and relationships between groups. In B B Smutts, D L Cheney, R M Seyfarth, R W Wrangham, and T T Struhsaker, editors, *Primate Societies*, pages 267–281. University of Chicago Press, Chicago, 1987.

[Cohen and Levesque, 1991] P. R. Cohen and H. J. Levesque. Teamwork. *Noûs*, 25(4):487–512, 1991.

[Conradt and Roper, 2003] L. Conradt and T. J. Roper. Group decision-making in animals. *Nature*, 421:155–158, January 2003.

[Dunbar, 1996] R. I. M. Dunbar. *Grooming, gossip and the evolution of language*. Faber and Faber, London, 1996.

[Grosz and Sidner, 1990] B. J. Grosz and C. L. Sidner. *Intentions in Communication*, chapter Plans for discourse, pages 417–445. MIT Press, Cambridge MA, 1990.

[Hemelrijk, 2002] C. J. Hemelrijk. Self-organizing properties of primate social behavior: A hypothesis for intersexual rank overlap in chimpanzees and bonobos. *Evolutionary Anthropology*, Suppl 1:91–94, 2002.

[Hill and Dunbar, 2002] R. A. Hill and R. I. M. Dunbar. Climatic determinants of diet and foraging behaviour in baboons. *Evolutionary Ecology*, 16:579–593, 2002.

[Hill *et al.*, 2003] R. A. Hill, L. Barrett, D. Gaynor, T. Weingrill, P. Dixon, H. Payne, and S. P. Henzi. Day length, latitude and behavioural (in)flexibility in baboons. *Behavioral Ecology and Sociobiology*, 53:278–286, 2003.

[Hill, 1999] R. A. Hill. *Ecological and demographic determinants of time budgets in baboons: Implications for cross-populational models of baboon socioecology*. PhD thesis, University of Liverpool, 1999.

[Jerison, 1973] H. J. Jerison. *Evolution of the brain and intelligence*. Academic Press, New York, 1973.

[Jolly, 2001] C. J. Jolly. A proper study for mankind: Analogies from the papionin monkeys and their implications for human evolution. *Yearbook of Physical Anthropology*, 44:177–204, 2001.

[Lomnicki, 1992] A. Lomnicki. Population ecology from the individual perspective. In D. L. DeAngelis and L. J. Gross, editors, *Individual-based models and approaches in ecology*, pages 3–17. Chapman and Hall, New York, 1992.

[Maynard-Smith, 1995] J. Maynard-Smith. Life at the edge of chaos. *New York Rev. Books*, March 2:28–30, 1995.

[Milinski and Parker, 1991] M. Milinski and G. A. Parker. Competition for resources. In J R Krebs and N B Davies, editors, *Behavioural Ecology*, pages 137–168. Blackwell, Oxford, 1991.

[Mithen, 1994] S. J. Mithen. Simulating prehistoric hunter-gatherer societies. In N Gilbert and J Doran, editors, *Simulating Societies: the Computer Simulation of Social Phenomena*, pages 165–193. UCL Press, London, 1994.

[Reynolds, 1987] Craig W. Reynolds. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, pages 25–34. ACM Press, 1987.

[Richard, 1985] A. F. Richard. *Primates in Nature*. W.H. Freeman and Company, New York, 1985.

[Robbins and Robbins, 2004] Martha M. Robbins and Andrew M. Robbins. Simulation of the population dynamics and social structure of the virunga mountain gorillas. *Americal Journal of Primatology*, 63(4):201–223, 2004.

[Stacey, 1986] P. B. Stacey. Group size and foraging efficiency in yellow baboons. *Behavioral Ecology and Sociobiology*, 18:175–187, 1986.

[Tambe, 1997] M. Tambe. Towards flexible teamwork. *Journal of Artificial Intelligence Research*, 7:83–124, 1997.

[Tucker, 1970] V. Tucker. The energetic cost of locomotion in animals. *Comparative Biochemistry and Physiology*, 34:841–846, 1970.

Tolerance and Sexual Attraction in Despotic Societies: A Replication and Analysis of Hemelrijk (2002)

Hagen Lehmann and **JingJing Wang** and **Joanna J. Bryson**
University of Bath
Artificial models of natural Intelligence
Bath, BA2 7AY United Kingdom
h.lehmann@cs.bath.ac.uk, ext-jingjing.3.wang@nokia.com, j.j.bryson@cs.bath.ac.uk

Abstract

Most primate societies are characterised by hierarchical dominance structures. Males are usually dominant over females, but in periods of sexual attraction (during females period of tumescence) male 'tolerance' towards females rises. (Hemelrijk, 2002) shows in a model that this 'tolerance' is created as a side effect due to the rise of female dominance during periods of sexual attraction. This rise is in turn the consequence of the more frequent approaches of males towards females during these periods. In Hemelrijk's model the males gain no benefit from 'tolerating' females and they only do so at high aggression levels as a kind of 'respectful timidity', because some of the females have become dominant over them.

This paper replicates and examines the results of Hemelrijk's study. We have found that some of Hemelrijk's results are highly reliant on aspects of the model that are not well supported by the current primate literature. We analyse the mechanisms underlying her results, and suggest data that should be sought from observation logs of real primate colonies that would support or overturn the model.

1 Introduction

In this paper, we examine the best-established AI model of primate social systems, Hemelrijk's DomWorld (Hemelrijk, 1999b,a, 2000, 2002). Hemelrijk models a large amount of primate behaviour using an incredibly simple model of social interactions based on spatial locations. In this paper, we replicate DomWorld, which allows us to examine the mechanisms underlying the system. We pay particular attention to the results from Hemelrijk (2002), the explanation of the increase of male tolerance experienced by females when they are sexually receptive (in *tumescence*). This particular experiment, situated in a wider model of differences between species in classifications of primate social structures, gives us a great deal of insight into the validity of Hemelrijk's approach.

We begin this paper by describing the primate social data to be explained and then by reviewing Hemelrijk's contributions. We then present our replication and our initial insights into the working of the DomWorld mechanisms. Finally, we discuss the validity of the model and propose specific data to look for that will either support or undermine the DomWorld model.

2 Background

Most primate species are highly social. They live in structured societies which can be characterised as having more or less steep dominance hierarchies. A steep hierarchy is one in which individuals would never consider violating rank, for example a lower-ranked individual would not take any food in the presence of a higher ranked individual. In a more shallow hierarchy, dominant animals show greater tolerance of subordinate behaviour, and considerations of rank plays less of a role in ordinary action selection. The difference between these social structures have been most studied in macaque societies (see for a recent review Thierry et al., 2004). Societies characterised by steep hierarchies are often referred to colloquially as *despotic*, while those with the less rigid dominance structures are called *egalitarian*. When a dominant animal allows subordinate animals to take advantage of resources in its presence, the dominant animal is said to be expressing *tolerance*.

Tolerance is considered one of the most basic forms of conflict resolution (de Waal and Luttrell, 1989). It might be difficult to see tolerance as an action to be selected, since it seems more like a form of inaction. However, if an agent is very inclined to preserve resources (including its own social rank), then expressing tolerance can require considerable inhibition of strong inclinations. In some species, for example, this is achieved by the apparently deliberate averting of gaze or even moving away from a resource in order to avoid witnessing a desired event, such as allowing a juvenile throwing a tantrum to feed. This shift in visual attention is necessary if witnessing such an event would automatically trigger an emotional / species-typical response that would in turn prevent the completion of the feeding.

The structure of a primate society is also correlated with a number of other characteristics (de Waal and Luttrell, 1989; Thierry, 2000; Hemelrijk, 2002). Societies that are more despotic also tend to have more violent or aggressive interactions. On the other hand, their tend to be fewer conflicts than in egalitarian societies. In egalitarian societies, there are more frequent conflict interactions, but many of these involve

no injury or violent dispute. They may for example involve only hissing or snatching.

In most primate hierarchies males are usually dominant over females, due to their greater size, strength and aggression. However, during the female sexually attractive period of tumescence, chimpanzee males, for instance, allow females priority in food access (Yerkes, 1940). This has been explained as a probably cognitive strategy — an exchange for copulation — which is adaptive in that it also therefore produces offspring (Goodall, 1986; de Waal and Luttrell, 1989; Stanford, 1996).

Hemelrijk and her colleagues have proposed a cognitively-minimalist explanation of this change in behaviour. Hemelrijk claims that there is no statistical evidence for such exchanges for food (Hemelrijk et al., 1992), neither is there any increase in related offspring (Hemelrijk et al., 1999). Hemelrijk (2002) demonstrates a model where such a change in dominance occurs in despotic societies even without any benefit for the males, but as a simple consequence of the higher frequency of dominance interactions between the sexes brought on by the male's attraction to the females.

Hemelrijk claims that in her models, under the condition of high aggression intensities, males show tolerance towards females. Her evidence of tolerance is that, in her model, in times of sexual attraction, females may achieve ranks higher than males, while in other times they do not. Females are modelled as initially 50% weaker than males, and are persistently 20% less aggressive, which explains why this such outcomes are improbable in general. However, once an animal achieves a higher rank, their power is assumed (in these models) to also increase.

Hemelrijk explains her findings as a side effect of the higher frequency with which males approach the females. Normally, animals tend to avoid invading each other's 'personal space' and triggering a conflict unless they are of a higher rank than the animal they are approaching. However, in times of sexual attraction, Hemelrijk's males ignore rank in approaching females. Further, in Hemelrijk's model, the outcome of a dominance interaction is highly influenced by the extent to which it was unexpected. Thus if a very low ranking female happens to win a competition (which there is always a small chance of success — the probability being inversely proportional to the discrepancy in rank) then she will suddenly achieve a much higher rank.

Consequently, the opportunity for a low ranking female to win an interaction will rise as more males approach her. Thus she could become more dominant then some of the males, who will nonetheless continue approaching her, consequently likely increasing her rank as they fail in their subsequent dominance disputes. Therefor this 'tolerance' is more a 'respectful timidity' towards higher ranking females. The males will approach but not attack simply because she has a higher rank.

Thus a behaviour typically described as complex or even cognitive could, according to Hemelrijk's model, arise without any corresponding cognition. This change could be introduced to the species through a single exogenous factor, such as the availability of food resources, if this leads to an increase in aggression. This higher aggression then leads to a more despotic society in which in the periods of sexual attraction the dominance of the females rises as shown in the model and explained above.

Many researchers have expressed skepticism about Hemelrijk's work because of her anti-cognitivist stance. People who work closely with apes feel that it is 'obvious' that the animals have some cognitive capacity, or at least that when humans express very similar behaviour, they subsequently report having cognitive state.

Because we were curious about Hemelrijk's model and wished to understand it better, and because no version of DomWorld is freely available online, we have replicated Hemelrijk's work. In so doing we were able to examine the assumptions behind the model, and find out what aspects of the model were critical to its success in replicating primate behaviour.

3 Methods

Hemelrijk's model consists of a small troop of chimpanzees living near each other and occasionally having aggressive interactions, which result in shifts in dominance rank. After the model has run for a while quantitative descriptions of the agents' relationships are taken, such as the steepness of the dominance ranking hierarchy or the average centrality of an agent within its troop. These measurements are then compared to measurements made of real chimpanzees in natural situations to judge the quality of the model as a hypothesis of their behaviour.

3.1 The Model World

Our simulation was based on the model described by Hemelrijk (2002). She wrote her version in Object-Pascal and Borland Pascal 7.0. We used NetLogo 2.1, because, as a purpose-built modelling tool, it provides a relatively easy high-level language for quickly constructing models and visualising results. The world in which the agents interact is wrapped around on all sides and therefore resembles the geometrical structure of a torus. This is to avoid border effects and enable the agents to move in every direction. As described by Hemelrijk this space is of a size 200 x 200 units. It is a continuous space — agents have real-valued locations and can move in any of 360 directions. When an experiment starts, the agents set initially at random locations within a 30 x 30 parcel of this space. Each agent has a forward vision angle of 120 degrees (that is, it 'sees' or attends to agents that are 60 degrees to either side of its direction of forward motion), and a maximum perception range (*MaxView*) of 50 units. Consequently, at the beginning of the simulation, each agent will need to do no more than turn around to see all the other agents in the simulation. The visual limits restrict the amount of things that the agent is likely to attend to at any particular time.

Agent motion and social interaction is determined by a number of additional threshold parameters:

- a near-perception range, *NearView* of 24 units. Agents feel comfortable so long as they see some other agent within this range. If they do not, but they do see an agent (that is, one is within MaxView) then they will go towards that agent.

- a personal space parameter, *PerSpace*, of 2 units. Agents within this range of each other will have a dominance interaction.
- a search angle of 90 degrees. Agents rotate this amount if they can see no one within their MaxView.
- a waiting period. After an moves around or engages in a dominance interaction, it is assigned a random waiting time before it performs its next action.

The waiting period simulates foraging or resting in the wild — constant dominance interactions are not only unnatural but also make the troop so chaotic that spatial measurements of troop coherence and rank have no meaning. The waiting period is abbreviated when the agent observed a dominance interaction within its NearView. This is in accordance to observations in real animals, since in primate groups nearby fights are likely to trigger active behaviour in individuals (Galef, 1988).

In our experience, the model does not appear overly sensitive to most of the parameter values, although at the same time none of them can be eliminated and still maintain the action-selection model. However, the mode *is* particularly sensitive to the organisation of the waiting period. This is because many dominance interactions wouldn't happen if the relatively subordinate animal were able to avoid the relatively dominant one, but because only one animal tends to be moving at a time, the dominant one can invade the personal space of the subordinate.

In the simulations dealing with the impact of female tumescence on their dominance ranking there is one additional parameter *attraction* which is either *on*, indicating that all the females are tumescent, or *off*, indicating that none of them are.

3.2 The Interaction Structure

The interactions in the model are classified into two groups, one class consists of grouping interactions the other of dominance interactions. These two classes resemble the two forces which in nature on one hand drive groups apart and on the other hold them together in order to stabilise them (c.f. Reynolds, 1987).

For the grouping interactions Hemelrijk gives a set of four rules:

1. An agent which observes another agent within its personal space may perform a dominance interaction, depending on its own rank and the rank of the other agent. For such an interaction, first the nearest potential opponent is chosen. After an interaction, the winning agent moves one unit towards its opponent, while the loser turns around 180 degrees, plus or minus an angle drawn randomly from 45 degrees, then moves two units away.
2. If the agent detects nobody in its personal space, but can see other agents within its NearView, then — in trials without attraction — it moves one unit forward on its present course. In the attraction condition, if a VirtualMale can see a VirtualFemale, they will change their direction towards the nearest visible VirtualFemales and then move one unit forward.
3. If the agent detects no other agents within NearView, but there are agents within its MaxView range, then it changes direction toward the nearest one and moves one unit towards it.
4. If there are no other agents within MaxView, the agent turns in a search angle of 90 degrees at random to the right or left.

The dynamics of the simulation are such that, for any agent, there will always be at least one agent still in MaxView in some direction. Occasionally the troop splits, but the agents always reunites shortly. Given the rate of motion of the troop, the maximum duration of the waiting period, and the large difference between MaxView and NearView, no single individual can become "lost" from the troop.

In nature, dominance interactions between primates are characterised by the competition for resources such as food or potential mates. In order to gain stable access to such resources the different individuals within a group try to establish a rank in hierarchy that is as high as possible. This is achieved by constant interaction, which Hemelrijk calls in her paper a "long-term 'power' struggle." In the model there are no resources specified and the only trigger for interactions is spatial distance. The agents start 'fighting' when another agent is within their personal distance and the rank of the other is lower or equal to their own rank. The agent 'estimates' its chances to win, and if its chances seem good, then it engages in the competition (see below.)

Since the dominance values *within* each sex is equal at the beginning of a simulation, the outcome of every single interaction influences the chances of winning the next one. Such a system is self-reinforcing and has been shown empirically in many animal species (Hemelrijk, 2000).

The formula for determining the outcome of a dominance interaction was modelled after Hogeweg (1988) and Hemelrijk (1999b). Each agent has a certain dominance value, which is readjusted after every 'fight' the agent gets involved in. We called this value *Dom* according to Hemelrijk's notation. This variable is correlated both to the agent's rank and its ability to win an interaction. If one agent finds another agent in its PerSpace, it compares its own Dom-value with the Dom-value of the other. If its own value is higher or equal to the other it 'estimates' it has good chances to win and will therefore interact. The outcome of the interaction is calculated it with the following formula (from Hemelrijk, 2002, p. 734)

$$w_i = \begin{cases} 1 & \frac{Dom_i}{Dom_i + Dom_j} > Random(0,1) \\ 0 & else \end{cases} \quad (1)$$

Where Random(0,1) produces a random real value between 0 and 1.

In this calculation, w_i is the value which determines whether agent i has lost or won. Here 1 means victory and 0 defeat. The relative dominance value is compared with a randomly drawn number between 0 and 1. If if is greater then the drawn number, the agent wins. This means that higher an agent's rank is relative to its opponent, the more likely the agent is to win.

After a dominance interaction, the dominance values of both agents are adjusted according to the outcome, using roughly the same information.

$$Dom_i = Dom_i + \left(w_i - \frac{Dom_i}{Dom_i + Dom_j}\right) * StepDom \quad (2)$$
$$Dom_j = Dom_j + \left(w_i - \frac{Dom_i}{Dom_i + Dom_j}\right) * StepDom$$

The only exception to the above equations is that the lowest possible Dom-value is set to 0.01 in order to keep the Dom-values positive.

Hemelrijk calls this system for determining dominance values a *damped positive feedback system*, since in the case of winning the dominance value of the higher ranking agent goes up only slightly, but if the lower ranked agent wins its dominance value undergoes a great change. This is intended to reflect the fact that it is very unlikely for a low ranking individual to win an interaction with a high ranking one. Thus ranking is not changed much by an expected outcome, but it changes greatly for an unexpected one.

The amount of rank shift is also affected by another value *StepDom*. This value Hemelrijk uses to represent the intensity of the 'aggression' (or violence) of the interaction, which she hypothesises also correlates to the impact the interaction has on ranking. She uses a high StepDom value to represent the level of aggression in 'despotic' species, and a low StepDom value to represent the level in egalitarian ones. Values for StepDom can vary from 0 to 1 but are held constant within any give simulation, since they are considered to be determined by species. Although Hemelrijk calls this value 'aggression', notice that it has no direct impact on the probability or outcome of an interaction (see Eq. 1). Rather, it's impact is only indirect through its long-term impact on the dominance values which do determine both whether and how well an agent fights.

Another important element for correlating Hemelrijk's models to the real world is understanding her *coefficient of variation of dominance values*. This coefficient indicates the average variation between dominance ranks of the individuals in the troop. Hemelrijk interprets this coefficient as an indication of how 'despotic' or egalitarian a society is. Her hypothesis is essentially that there isn't a qualitative difference in how monkeys in an egalitarian society treat their superiors vs. how those in a despotic one do, but rather that every agent will show an equal amount of respect for a troop-mate with twice its dominance value. Thus Hemelrijk represents a despotic society as one with an unambiguous / 'steep' dominance hierarchy, with a great difference in rank between individuals, and an egalitarian one as having relatively ambiguous rankings.

3.3 Experimental Set-Up

For our attempted replications, we used the parameter settings Hemelrijk uses in several studies (Hemelrijk, 1999a, 2000). We used 8 agents in a troop, four of each sex ($N = 8$). As explained earlier, each agent had an personal space of 2 ($PerSpace = 2$), a vision angle of 120 degrees, an maximum perception range of 50 units ($MaxView = 50$) and near-perception range of 24 units ($NearView = 24$). The search angle was 90 degrees, the fleeing distance was 2 units ($fleeD = 2$), the fleeing angle was 45 degrees at random direction away from the opponent and the chasing distance was 1 unit ($chased = 1$) in the direction of the opponent.

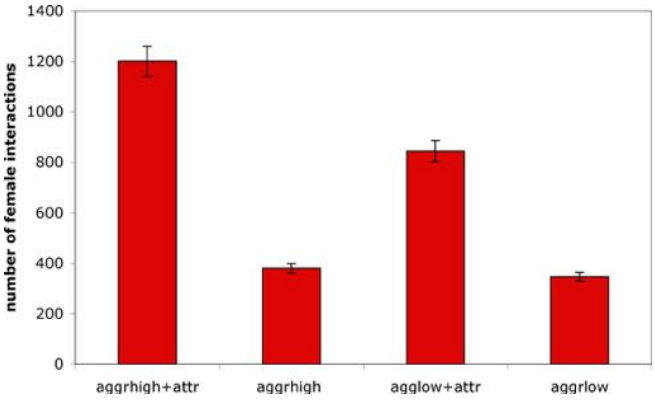


Figure 1: Total number of female interactions in different conditions. *aggrhigh+attr* = high aggression + attraction; *aggrhigh* = aggression high + no attraction; *agglow+low* = aggression low + attraction; *aggrlow* = aggression low + no attraction.

To resemble the difference in physical strength between males and females both sexes started out with different winning or loosing tendencies — that is the DomValues of females were half that of males ($virtual\ females = 8$, $virtual\ males = 16$). Also, females have only 80% of the aggression intensity (*StepDom*) of males. The experiment was conducted with 4 different conditions. We used two level of aggression to correlate with the two types of social interactions witnessed in different primate species. In the high level the StepDom value of males was 1 and of females 0.8, in the low aggression level the StepDom value of males was 0.1 and of females 0.08. These two aggression conditions were each run under two conditions of *sexual attraction* (either turned on or off) 10 times each, resulting in a total number of 40 runs. Each run was 42800 time units long.

4 Results

Our results match Hemelrijk's results to the extent that we used the same analysis, which we largely did in order to test the replication. The first figure shows a comparison between the number of interactions performed by virtual females during the different conditions. In the graph the total number of aggressive interactions initiated by virtual females is compared for all four different conditions used in the experiment.

In Figure 1 we can see that the number of virtual female dominance interactions increases significantly in conditions with sexual attraction in both intensities of aggression (Mann-Whitney, N = 10, U = 0, $p < .001$, two-tailed, Mann-Whitney, N = 10, U = 0, $p < .001$, two-tailed). That means females are involved in considerably more interactions when they are attractive. The aggression level amplifies the result, even though this effect for the aggression is rather weak

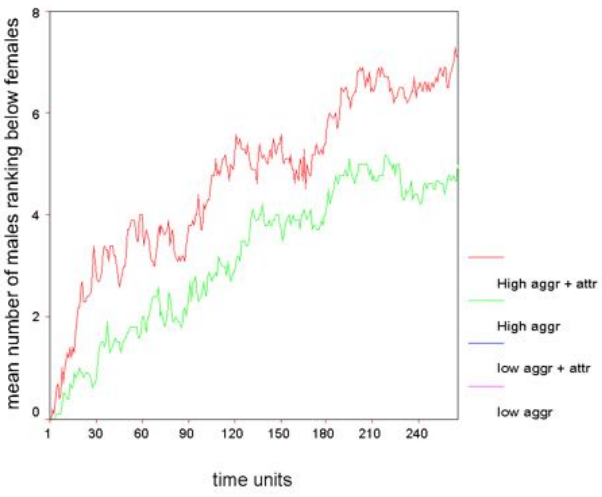


Figure 2: The dominance of virtual females as the sum of the number of males ranked below each female at different times in different conditions.

(Mann-Whitney U-Test, N = 10, U = 24 $p < .049$, two-tailed).

Figure 2 shows the dominance of virtual females as the sum of the number of males ranked below each female at different times in different conditions. We can see that, as reported in Hemelrijk, the female dominance in conditions with high aggression level increase over the time, but that they stay constant in conditions with a low aggression level.

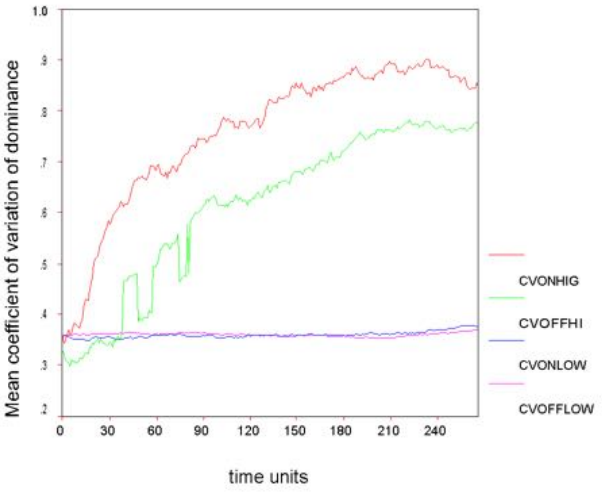


Figure 3: Distribution of the coefficient of variation of dominance values in different conditions for both sexes.

Figure 3 is the classic Hemelrijk result. It shows the distribution of the coefficient of variation of dominance values for both sexes (see discussion in previous section.) If the aggression is high, there will be a steeper hierarchy — the difference between rank values will be larger. This is true both within and between sexes. Attraction amplifies this result, despite the fact that some females may outrank some males in this condition.

The last two figures show the change of dominance values for both sexes in conditions with high and with low levels of aggression. With high aggression a constant change in the dominance structure is noticeable greater and greater differentiation / steepness in the hierarchy. With low aggression there is only very little change in the dominance values. This creates a very stable hierarchy where the females never gain a higher positions in the group.

The conclusion of these results is, that only in groups with a high level of aggression females are able to gain higher positions in the social hierarchy. The attraction amplifies this effect, but plays a secondary role.

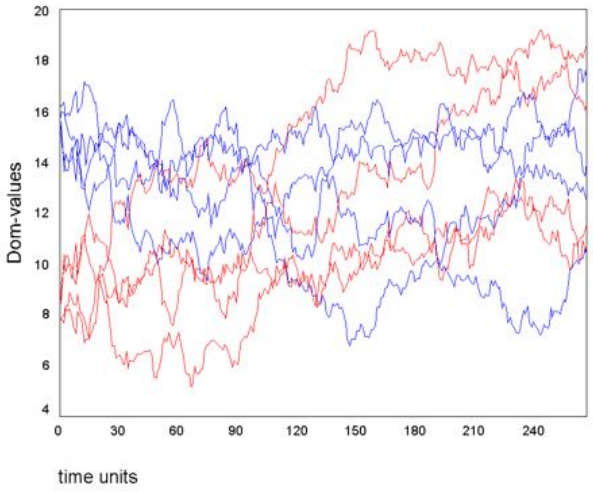
5 Discussion

Our results show the same structure as the results in the original study by (Hemelrijk, 2002) Figure 3A, p. 739 Figures 4A B C, p. 741) and can therefor be seen as a replication. In general, the diversity of different dominance values between individuals increases if there is a high aggression level existing within the population. In conditions with low aggression levels this effect does not appear, even though the results in this model show that the increase of interactions between virtual females and virtual males depends not on the increased level of aggression but on the existence of female attraction. In this first result, we can see that the level of aggression has no (or at best only very little) influence on the number of interactions between the individuals, yet in both conditions with female attraction the increase of interactions is significant.

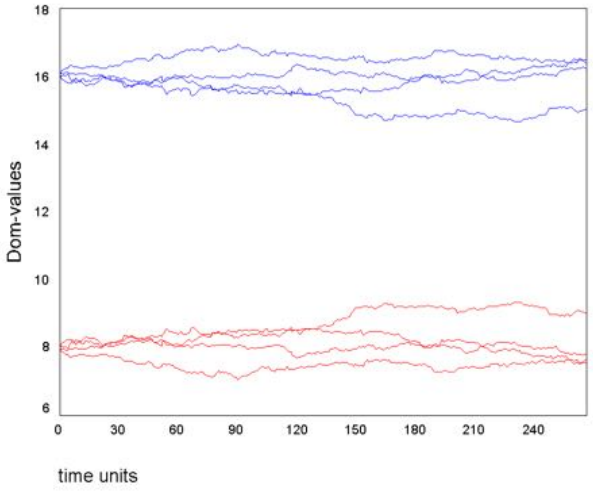
The most interesting effect is the change in dominance values towards more dominant females and as a possible consequence a change in group structure. This connection between higher interaction frequency and the dominance value change Hemelrijk claims in her article (p. 742) could be a simple explanation for the observed natural phenomenon of male tolerance towards females in their period of sexual attractiveness. Given our understanding of Hemelrijk's model derived from our replication, we will now examine these claims more closely.

One of the strengths of agent-based modelling (ABM) is its ability to demonstrate whether theories of the origin of behaviour can be explained by a given model of how an agent selects its actions. In particular, as with the rest of science, there is an emphasis in ABM on looking for the simplest possible explanation that fits the data. We look for the origins of complex behavioural patterns on a social level as emergent from simple behaviour in the individual.

We need to realize though, that this is not only a following of the principle of parsimony for reasons of the philosophy of science. It may also be a case of looking for our keys under the light of the street lamp rather than over in the dark where we lost them. Complex individual behaviour is difficult to program, takes a long time to execute in simulation, and then is difficult to analyse. So we may have a strong bias towards looking for overly simple solutions. Thus while on the one hand we need to be open-minded and be sure to understand



(a) High Level Aggression



(b) Low Level Aggression

Figure 4: Distribution of dominance values at a high level and at a low level of aggression. In both conditions, the males start off initially higher than the females.

correlations where we find them, on the other hand we cannot allow our biases to blind us to a situation where data may not fit the predictions of our model. Guarding against this bias is just as important as guarding against its opposite, the overly-cognitive explanations.

The Hemelrijk model we have replicated seems to be a good analogue system for macaque behaviour. Her Dom-World model shows that apparently complex behaviours in primate societies like 'male tolerance' or 'female assertiveness' can be created in computer-generated primate societies with only a few simple assumptions about individual behaviours. The effect of female dominance appears for example in the conditions with high aggression and is consolidated by a high level of attractiveness in the females. Hemelrijk notes the difference between this and the classical explanations for this phenomenon, which propose exchanges involving food for sexual opportunities (Goodall, 1986). Hemelrijk's model does not include any food or sex, yet still leads to analogous results.

Now that we have a working model, we can try to understand exactly where and how these phenomena 'emerge'. We can now analyse what the critical factors of the model are, and look for biological correlates that would either prove or disprove the model.

The effect of the model is based on two major assumptions:

1. the self-reinforcing effect of domination, and
2. the fact that females attract males in their time of tumescence, but that males are not attractive to females.

The first assumption relates to the fact that the dominance value DOM of an individual i (operationalised as the ability to win a fight) increases with a victory and decreases with a defeat. Although this self-reinforcement is a well-known phenomena that has been studied extensively in laboratory animals such as mice, we are somewhat skeptical of the exact extent to which this model depends on these factors. In Hemelrijk's model, the strength of the effect is determined by the dominance ranking of the opponent, the 'level of aggression' (that is, the step-value assigned to this species) and chance. The result of a fight is calculated with Equation 1 repeated here:

$$w_i = \begin{cases} 1 & \frac{Dom_i}{Dom_i + Dom_j} > Random(0,1) \\ 0 & else \end{cases} \quad (3)$$

Again as a reminder, the dominance level after a fight is calculated with Equation 2:

$$Dom_i = Dom_i + \left(w_i - \frac{Dom_i}{Dom_i + Dom_j} \right) * StepDom$$
$$Dom_j = Dom_j + \left(w_i - \frac{Dom_i}{Dom_i + Dom_j} \right) * StepDom$$

As we emphasised earlier, Hemelrijk has defined the factor $StepDom$ to mean aggression. An individual therefore increases its ability to win a fight (its dominance) most, if it wins against an individual with a preferably much higher

dominance level and if the aggression level in the group is high.

The aggression is therefor the crucial value which decides within the system how far an individual can go up or fall down in the hierarchy as the result of a single fight. This is largely the basis of the reinforcement effect of domination, but to what extent does this effect exist in nature? Hemelrijk's text only mentions observations on bumblebees and other computational models as examples (p. 743 f). Thinking about it in a intuitive way it might be plausible, that self-confidence about winning a fight increases, if one wins against someone much stronger. Further, we know that even in adult mammals, growth hormones can be triggered by success in social competitions. Nevertheless in a real fight the body size and strength is at least as important as the psychological status of the individual.

To test the validity of Hemelrijk's model, we need to use the documented history of dominance hierarchies in real animals. We would need to look carefully at the relatively rare events where a lower-ranked animal bested a higher ranking animal, and see what the impact is on the troops dominance structure before and after. We should look in particular for the following factors:

- If one agent defeats another that vastly outranks it in a dominance interaction, do the two agents immediately change ranks within the troop? In other words, is an unexpected outcome from a fight likely to have a very significant effect? If this is true, it would validate the use of relative dominance values in Equation 2.
- In comparing across species, does it take fewer interactions to advance rank in a 'despotic' species? If this is true, then it would justify the use of StepDom in Equation 2.
- Within species, if a fight is more violent (e.g. if blood is drawn compared to mild beating, or if there is mild beating compared to a non-physical interaction) does it have more impact on dominance hierarchy? If this is so, then it makes sense to refer to StepDom as 'aggression' and it would further validate it's use in Equation 2.
- Are females more likely to engage in fights when they are in tumescent? If not then this model cannot account for their increased dominance.
- Do females only become dominant during their tumescence in 'despotic' species? Given that the prime indication in Hemelrijk's model of increased dominance for the females is the males' increased tolerance of them, discriminating an increase of rank in an egalitarian species may be difficult, since these species are definitionally tolerant towards all group members. But it is a predicition of the model.
- Is it true that when an animal in an egalitarian species is *clearly* outranked by another animal, that those two animals' interactions will be similar to two more nearly ranked animals in a less egalitarian species? Or *is* there a qualitative difference in how different species behave with respect to dominance hierarchies? The answer to this question will serve to validate whether steepness of the dominance hierarchy is a good representation of despotism / egalitarianism.

Of course, this is complicated by the fact that establishing a dominance hierarchy is never easy — it's not clear that every animal will agree on the current hierarchy, and indeed some animals will behave differently with respect to others depending on what other animals are present (Harcourt, 1992). However, many groups work diligently to attempt to establish these sorts of records, so we can hope to test these predictions.

We need to also look critically at the second basic assumption, the idea that the female primates attract male primate in their fertile days. This is obviously true, but sexual attraction is bidirectional and therefor influences the grouping behaviour of females as well. Of course, it is possible that the male attraction is strong enough to overwhelm the data, or even that just putting high male attraction is a good approximation for mutual attraction. However, the question remains as to whether the mechanism exploited by the model — increased conflict leading to a higher probability of an occasional lucky win by the female that immediately catapults her high into the dominance hierarchy — is at all plausible.

6 Conclusions

We have presented a replication of Hemelrijk (2002) and an analysis of how her model works. We have also presented a critical list of suggestions for testing the validity of the mechanism. We suspect that the rules for determining dominance from the outcome of dominance battles are not sufficiently realistic and cannot fully explain the change in female dominance rank on their own. If we are right, then this model may need additional factors to explain this phenomena, possibly including cognitive state sufficient for the traditional theories of reciprocation.

Acknowledgements

Several MSc and undergraduate students over the years have explored using NetLogo in the AmonI group for this task. We owe particular gratitude to Wang (2003). All software used in this paper is available on request or from our website.

References

de Waal, F. B. M. and Luttrell, L. (1989). Toward a comparative socioecology of the genus *macaca*: Different dominance styles in rhesus and stumptailed macaques. *American Journal of Primatology*, 19:83–109.

Galef, B. G. J. (1988). Imitation in animals: history, definition, and interpretation of data from the psychological laboratory. In Zentall, T. and Galef, B. G. J., editors, *Social Learning: Psychological and Biological Perspectives*, pages 3–25. Erlbaum, Hillsdale NY.

Goodall, J. (1986). *The Chimpanzees of Gombe: Patterns of Behavior*. Harvard University Press.

Harcourt, A. H. (1992). Coalitions and alliances: Are primates more complex than non-primates? In Harcourt,

A. H. and de Waal, F. B. M., editors, *Coalitions and Alliances in Humans and Other Animals*, chapter 16, pages 445–472. Oxford.

Hemelrijk, C. K. (1999a). Effects of cohesiveness on intersexual dominance relationships and spatial structure among group-living virtual entities. In Floreano, D., Nicoud, J.-D., and Mondada, F., editors, *Proceedings of the Fifth European Conference on Artificial Life (ECAL99)*. Springer.

Hemelrijk, C. K. (1999b). An individual-oriented model on the emergence of despotic and egalitarian societies. *Proceedings of the Royal Society London B: Biological Sciences*, 266:361–369.

Hemelrijk, C. K. (2000). Towards the integration of social dominance and spatial structure. *Animal Behaviour*, 59(5):1035–1048.

Hemelrijk, C. K. (2002). Self-organization and natural selection in the evolution of complex despotic societies. *Biological Bulletin*, 202(3):283–288.

Hemelrijk, C. K., Meier, C. M., and Martin, R. D. (1999). 'friendship' for fitness in chimpanzees? *Animal Behaviour*, 58:1223–1229.

Hemelrijk, C. K., van Laere, G. J., and van Hooff, J. A. R. A. M. (1992). Sexual exchange relationships in captive chimpanzees. *Behavioural Ecology and Sociobiology*, 30:269–275.

Hogeweg, P. (1988). MIRROR beyond MIRROR: Puddles of LIFE. In *Artificial life: Santa Fe Institute studies in the sciences of complexity*, pages 297–316. Addison-Wesley, Reading, Massachusetts.

Reynolds, C. W. (1987). Flocks, herds, and schools: A distributed behavioral model. *Computer Graphics*, 21(4):25–34.

Stanford, C. B. (1996). The hunting ecology of wild chimpanzees: Implications ofr the evolutionary ecology of pliocene hominids. *American Anthropologist*, 98:96–113.

Thierry, B. (2000). Covariation of conflict management patterns across macaque species. In Aureli, F. and de Waal, F. B. M., editors, *Natural Conflict Resolution*, chapter 6, pages 106–128. University of California Press.

Thierry, B., Singh, M., and Kaumanns, W., editors (2004). *Macaque Societies*. Cambridge University Press.

Wang, J. J. (2003). Sexual attraction and inter-sexual dominance among virtual agents — replication of hemelrijk's domworld model with netlogo. Master's thesis, University of Bath. Department of Computer Science.

Yerkes, R. M. (1940). Social behavior of chimpanzees: Dominance between mates in relation to sexual status. *Journal of Compartative Psychology*, 30:147–186.

Collective Action Selection in Social Insect Colonies

James A. R. Marshall
Department of Computer Science
University of Bristol, Merchant Venturers Building
Woodland Road, Bristol BS8 1UB, UK
marshall@cs.bris.ac.uk

Abstract

Action selection is a problem that is faced not only by individual organisms, but also by collectives of organisms. Colonies of social insects are highly integrated units that frequently perform optimal collective action selection in a decentralised manner. Social insect colonies provide very accessible model systems for studying the mechanisms underlying such action selection processes. This paper discusses models of two action selection mechanisms in insect colonies, and speculates as to the potential for comparing action selection in such colonies to action selection in individuals.

1 Introduction

Action selection by individual organisms is a well studied problem. However, groups of organisms must also frequently perform action selection. The action selection problem becomes acute for highly integrated units such as colonies of social insects, where consensus during collective action selection is critical for the continued survival of the colony. What makes the action selection problem hard for insect colonies is their decentralised nature. While colonies of ants or honeybees do typically have a queen, these are only nominally in charge of the colony's fate. Certainly there is no individual within the colony, whether the queen or another, who takes the crucial decisions on the colony's behalf and issues instructions for their implementation. Instead, the colony has a rather flat management-hierarchy, and decisions are reached in a decentralised manner via local interactions between colony members. As social insect colonies are very amenable to experimental manipulation and observation, far more so than primate brains for example, they provide excellent model systems to study how collective decision making can be realised. In this paper we describe two such systems, and models that have been constructed to explain their behaviour. Furthermore, similarities between insect colonies and populations of neurons suggest that lessons learned from the study of group decision making in social insects may also be applicable to individual decision making, and vice versa.

2 Nest Site Selection by *Temnothorax albipennis*

Colonies of the ant *Temnothorax albipennis* (formerly *Leptothorax albipennis*) need periodically to emigrate from their current nest site to a new one. Two typical reasons for such emigrations are that either the original nest site has been rendered uninhabitable, perhaps being destroyed by a larger animal, or the colony's numbers have grown such that the current nest site is no longer large enough. During emigration it is normally very important that the colony avoids being split between two or more nest sites; the colony is so tightly integrated that the queen, all other adult ants, and all brood items need to be in the same nest site to ensure the colony's survival. The details of this emigration process have been elucidated by experimental observation (Mallon *et al*., 2001, Pratt *et al*., 2002, Franks *et al*., 2003, Dornhaus *et al*., 2004). When an emigration begins, scout ants from the colony leave the original nest site and search for potential new nest sites in the vicinity. On finding a potential site, a scout will assess several criteria, such as internal area (Mallon & Franks, 2000), structural integrity, darkness, etc., and integrate these different criteria into a single quality measurement (Franks *et al*., 2003a). This measure translates into a time delay before recruitment that is inversely proportional to the perceived quality of the site. After delaying, the scout will recruit other scouts to assess the same site and in turn recruit others, thus providing a kind of multiple "second opinion". When a scout recruits to a potential site, she initially recruits via a slow process known as "tandem running", in which the scout leads another ant to the site, maintaining physical contact throughout. However, if a scout enters a potential site and discovers that it contains a sufficient number of ants from the same colony, she will change her subsequent recruitment mode to a process known as "social carrying". Social carrying involves the scout picking up another passive ant or brood item and carrying it; social carrying is approximately three times faster than tandem-running (Pratt *et al*., 2002). The number of nest mates that must be in a potential site to trigger social carrying is known as the "quorum threshold". The quorum threshold is a key control device in the colony's decision-making process, allowing the colony to achieve slow but accurate decisions,

or fast but inaccurate decisions, by having a high or low quorum threshold respectively (Franks *et al*., 2003b). Low quorum thresholds mean that scouts begin social-carrying earlier, leading to more individualistic decision-making, while high thresholds mean that scouts take longer to begin social carrying, allowing them to recruit many other scouts to give their verdict on a site's quality. Colonies respond adaptively to the urgency of their emigration by varying the quorum threshold appropriately. Colonies use low quorum thresholds and hence fast, inaccurate decisions when their original nest has been destroyed, or they find themselves in a harsh environment. However colonies use high quorum thresholds to achieve slower, more accurate decisions when there is no urgency, for example when the original nest site is intact and they are simply searching for a superior nest site (Dornhaus *et al*., 2004).

A simple ordinary differential equation (ODE) model, formulated by Pratt *et al*. (2002), is able to recreate this dependence of speed and accuracy of decision making on quorum threshold (figure 1).

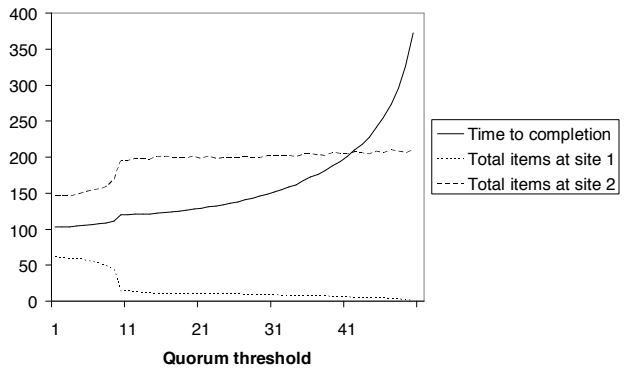


Figure 1. Output from a mathematical model of house-hunting in *T. albipennis* parameterised to replicate their speed-accuracy trade-off. The graph shows the decision time (time until the original nest site is empty in minutes) and accuracy (items ending up at site 2, out of 208 total) achieved with different quorum thresholds (results from the model presented in Pratt *et al*. (2002), figure reproduced from Marshall *et al*. (2005)).

Subsequent work by Pratt *et al*. (2005) resulted in a more realistic individual-based model, extensively validated against data from biological experiments, hence demonstrating the sufficiency of individual-based rules to achieve the collective behaviour observed in *T. albipennis*.

However, Marshall *et al*. (2005) revealed this original result to be an artefact of a very low rate of switching between alternative nest sites. Ant scouts are observed to be able to change the target of their assessment or recruitment efforts, if they discover an alternative site during an emigration. A fundamental assumption of the Pratt *et al*. ODE model was that such switching was unidirectional, from inferior to superior choice, based on experimental observations that ant scouts visiting both alternative nest sites subsequently confine their recruitment to the superior of the two. Marshall *et al*. noted that increasing this switching rate thus collapsed a two solution decision problem to a single solution decision problem, and even with a moderate increase in the switching rate the colony could achieve a completely accurate decision using the minimum quorum size, with no corresponding increase in decision time (figure 2). Marshall *et al*. constructed a similar individual-based model to Pratt *et al*. (2005), adding a variable time cost for nest site assessment, and a variable degree of noise, and found that both of these mitigated the benefits of increased preference switching rate. The conclusion of this investigation was thus that the time cost and noise that real colonies of *T. albipennis* experience when assessing nest sites probably mean that their switching rate is Pareto-optimal; increasing switching rate might improve accuracy but only at the cost of speed (Marshall *et al*., 2005).

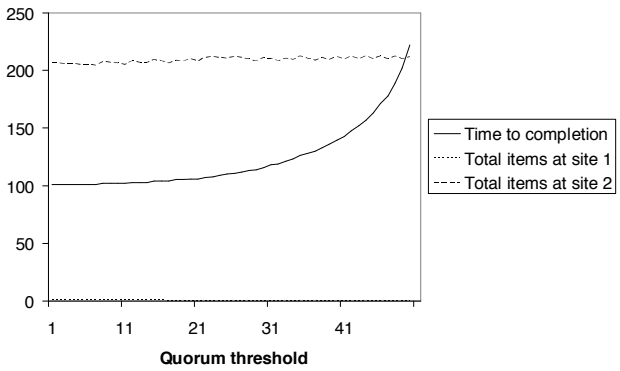


Figure 2. Output from the same mathematical model (Pratt *et al*., 2002) with the same parameters used to generate figure 1, but with the switching rate, ρ_{12}, from the inferior to superior nest site increased from 0.008 to 0.06. The result of this change is that the speed-accuracy trade-off observed in figure 1 disappears; minimal quorum size leads to perfectly accurate decisions without any corresponding increase in decision time, and decision time becomes less influenced by quorum threshold used (figure reproduced from Marshall *et al*., 2005).

3 Navigation During Swarming by *Apis mellifera*

During emigration, colonies of the honeybee *Apis mellifera* search for potential new nest sites in a similar way to *T. albipennis*. Scouts leave the hive, discover potential new nest sites, then return to the hive and advertise their location via the waggle-dance (von Frisch, 1967). Other scouts thus acquire information about the location of alternative nest sites, then go and evaluate them. Over time, scouts progressively cease dancing for nest sites, until there is only one nest site being advertised in the hive and a consensus is reached (Seeley, 2003). However, the recruitment method of *A. mellifera* poses a problem for implementation of the selected action. Unlike *T. albipennis*, where a population of transporters is built up who can transport passive nestmates and brood items to the new nest site, members of the honeybee colony must make it to the new site under their own power. Once consensus is reached, the colony swarms, and

must fly as a whole to the selected new nest site. Again, as with *T. albipennis*, the colony is so well integrated functionally that the colony's survival depends, in large part, on its ability to maintain cohesion and avoid being split between two or more nest sites. Thus, the problem is this: only a very small proportion of the colony have been actively scouting, and so know the way to the nest site, yet this minority of informed individuals must guide the entire colony there.

While the existence of this behaviour is well known, the precise mechanisms underlying it have for some time been the domain of speculation only. Couzin *et al*. (2005) have now investigated a model of collective action selection during navigation, applicable to *A. mellifera*, with interesting results. Specifically, Couzin *et al*. found that a small minority of informed individuals within the group could lead the entire group in the appropriate direction, without any individual knowledge about: which other individuals are informed, whether there are any other informed individuals at all, and whether there are other informed individuals with a different target. Couzin *et al*. found that for a group size of 200 (much smaller than a typical honeybee colony) only 5% of the population need be informed, and also found that the size of this required minority decreased as group size increased. Thus Couzin *et al*'s model explains how simple individual rules might allow a very small proportion of honeybee scouts to guide the entire swarm to their preferred nest site.

4 Prospects for Comparison of Collective and Individual Action Selection

A student of the collective action selection mechanisms employed by insect colonies might, at first, believe that they are entirely distinct from the mechanisms used by individuals. However, on further inspection there are intriguing similarities between the two. Collective action selection in an insect colony arises from the interactions between sub-populations of individual insects; in the same way, individual action selection in the brain arises from the interactions between sub-populations of neurons. Additionally, features of individual action selection, such as exploitation of speed-accuracy trade-offs (Edwards, 1965), can also be seen in some collective action selection mechanisms (Franks *et al*., 2003a). Some details may differ; for example, whereas a brain performing an action selection task such as saccading simultaneously integrates sensory information from a variety of sources, an insect colony must implement a sampling strategy to acquire information. Thus the insect colonies' action selection problem is actually more closely related to a bandit problem (Marshall *et al*., 2005). Nevertheless, there remains the tantalising prospect that similar mechanisms may underlie both collective and individual action selection. Thus, lessons learned from collective action selection may inform understanding of individual action selection, and vice versa.

Acknowledgements

I am grateful to colleagues in the Ant Lab, School of Biological Sciences, University of Bristol for the benefit of their expertise, and in particular Nigel Franks and Anna Dornhaus for their collaboration. I also thank Tim Kovacs for his collaboration, and Rafal Bogacz for fascinating discussions on action selection in the brain.

References

Couzin, I. D., Krause, J., Franks, N. R. & Levin, S. A. (2005) Effective leadership and decision-making in animal groups on the move. *Nature* **433**, 513-516.

Dornhaus, A., Franks, N. R., Hawkins, R. M. & Shere, H. N. S. (2004) Ants move to improve: colonies of *Leptothorax albipennis* emigrate whenever they find a superior nest site. *Anim. Behav*. **67**, 959-963.

Edwards, W. (1965) Optimal strategies for seeking information: models for statistics, choice reaction times, and human information processing. *J. Math. Psy*. **2**, 312-329.

Franks, N. R., Dornhaus, A., Fitzsimmons, J. P. & Stevens, M. (2003a) Speed vs. accuracy in collective decision making. *Proc. R. Soc. Lond. B* **270**, 2457-2463.

Franks N. R., Mallon E. B., Bray H. E., Hamilton M. J. & Mischler T.C. (2003b) Strategies for choosing between alternatives with different attributes: exemplified by house-hunting ants. *Anim. Behav*. **65**, 215-223.

Mallon, E. B. & Franks, N. R. (2000) Ants estimate area using Buffon's needle. *Proc. R. Soc. Lond. B* **267**, 765-770.

Mallon, E. B., Pratt, S. C. & Franks, N. R. (2001) Individual and collective decision-making during nest site selection by the ant *Leptothorax albipennis*. *Behav. Ecol. Sociobiol*. **50**, 352-359.

Marshall, J. A. R., Dornhaus, A., Franks, N. R. & Kovacs, T. (2005) Noise, cost and speed-accuracy trade-offs: decision making in a decentralised system. Under submission to *Interface: J. R. Soc*.

Seeley, T. D. (2003) Consensus building during nest-site selection in honey-bee swarms: the expiration of dissent. *Behav. Ecol. Sociobiol*. **53**, 417-424.

Pratt, S. C., Mallon, E. B., Sumpter, D. J. T. & Franks, N. R. (2002) Quorum sensing, recruitment, and collective decision-making during colony emigration by the ant *Leptothorax albipennis*. *Behav. Ecol. Sociobiol*. **52**, 117-127.

Pratt, S. C., Sumpter, D. J. T., Mallon, E. B. & Franks, N. R. (2005) An agent-based model of collective nest choice by the ant *Temnothorax albipennis*. To appear in *Anim. Behav*.

von Frisch, K. (1967) The Dance Language and Orientation of Bees. Belknap Press, Cambridge MA.

Visual Communication and Social Structure – The Group Predation of Lions

Alwyn Barry and Hugo Dalrymple-Smith
University of Bath
Department of Computer Science
Claverton Down, Bath, UK.
a.m.barry@bath.ac.uk

Abstract

[Creel, 1997] in a study of african hunting dogs suggested that, where the maximisation of net energy gain from hunting requires cooperation, cooperative hunting becomes an important part of the sociality of the hunting dogs. When considering Lion cooperative hunting [Scheel and Packer, 1991] suggested, in contrast, that Lions do not form groups to increase the intake of food for the group but for the individual, implying that cooperative behaviour in hunting has very little impact on the formation of prides. In using simulation to investigate the role of visual location in the group hunting behaviour of Lions it is shown that a minimal communication simulation can be derived if the dominance of pride members is taken into account. We conclude that agreed dominance permits the reduction of visual cues required to coordinate complex cooperative simulated behaviour.

1 Introduction

Cross-disciplinary engagement between behavioural biology and computer science has generated valuable insights and new directions for both biological and computer science research [Boden, 1996], such as in the understanding of flock formation [Reynolds, 1987] or fish schooling [Tu and Terzopoulos, 1994]. For example in [Zaera *et al*., 1996] biologists had proposed several different hypotheses regarding schooling behaviour, but these presented contradictory opinions about the reasons for schooling behaviour. Artificial life models of fish behaviour and schooling have provided new tools with which to evaluate the plausibility of biological hypotheses, and have provided computer scientists with valuable models of behaviour and control.

Considerable progress has been made in the modelling of aspects of behaviour in animals with relatively limited cognitive abilities (such as ants, spiders and fish). The modelling of higher animals is more problematic. With lower animals limited cognitive capabilities rule out consideration of complex models. The discovery of a simple mechanism to describe a behaviour in a higher animal does not rule out the use of a more complex mechanism by the animal, and the cognitive capabilities of the animal provide a large search space for plausible models. Nonetheless, the search for simple models of complex behaviour in higher animals is important not only for the Behavioural Biologist but also for the Computer Scientist. For example, in robotic control it is more likely that few robots with considerable computing power will be available. Behavioural algorithms based on the behaviour of tens or hundreds of simple animals will not necessarily scale down in useful ways with orders of magnitude fewer robots. In contrast, behavioural models derived from the smaller groups of higher animals may be a better fit.

[McFarland, 1994] distinguishes between 'eusocial' and 'cooperative' [1] behaviour where the latter is a behaviour selected intentionally by selfish agents to maximise individual utility in constrast to the former which is innate (genetically encoded). Whilst many large predators are lone hunters, some demonstrate selective behaviour – utilising lone hunting in certain environmental situations and small group cooperative hunting in others. However, there has been debate over when seemingly cooperative behaviours can be considered cooperative or are merely an extension of the selfish behaviour of agents forced by circumstance to be part of a group.

In their study of the hunting behaviour of Serengeti lions, [Scheel and Packer, 1991] noted that when cooperative behaviour does happen it appears more likely to occur in situations when the prey is larger, more difficult to kill or in long distance hunts. Using success of the hunt as a criteria, they suggest that lions do not form groups to increase the intake of food for the group but for the individual, implying that cooperative hunting does not exist but rather that opportunistic hunting is being displayed. The observation that amongst the lions there are some which take a less active role in hunting, a behaviour they define "cheating", is used to support this hypothesis.

From a study of African hunting dogs [Creel, 1996; 1997] have suggested that cooperative hunting should not be judged on the success of the hunts, but on the food intake per day against the energy spent during the hunt. Using this criteria, [Creel, 1997] was able to show that the packs formed by the African hunting dogs were optimal for pack sizes of 8 – 11. They suggest that cooperative hunting plays an important part

[1] [Cao *et al*., 1994] defines "cooperative" behaviour as follows: 'a [multi-agent] system displays cooperative behaviour if, due to some underlying mechanism, there is an increase in the total utility of the system'.

in the sociality of african hunting dogs. [Creel, 1997] extends this argument to prides of lions. [Griffin, 1984] suggests that there must be some form of conscious decision making behind the hunting behaviour of the lions. Although this opinion is largely based upon on a single observation, he argues that it is difficult to believe that with the success that cooperative hunting brings the lions are not at some level aware of the benefits of planning such attacks. In contrast [Scheel and Packer, 1991] seem to hold the position that cooperative hunting has little impact on pride formation. [Schaller, 1972; Stander, 1991] do not commit to such an opinion. Stander notes that the question exists, and that the research simply shows the benefits of cooperative hunting. Schaller is satisfied with a statement of two possible positions, either that pre-planned cooperative hunting is taking place or that the lions are simply making use of the opportunities afforded by the presence of other lions.

We have conducted a preliminary investigation of these claims using a simplified simulation of group predation. From a review of lions' group hunting behaviour (see section 2) we hypothesised that the use of attraction / repulsion dynamics would lead to emergent group dynamics that simulates this behaviour. Our results demonstrate that to achieve the observed behaviour in simulation the addition of a dominance hierarchy between the simulated predators is required.

2 The Predators and their Prey

There are surprisingly few academic sources on the hunting methods used by lions, but those that are available provide a useful level of detail. The research available focuses on two different lion societies; from the Etosha National Park [Stander, 1991] and from the Serengeti National Park [Schaller, 1972].

2.1 Prey

Lions from the Etosha National Park were observed to have hunted and killed 16 different species of prey. Three species accounted for 83% of the hunting activity . . . springbok, zebra and wildebeest. The Serengeti lions were mainly observed hunting warthog, zebra, gazelle wildebeest and buffalo. This slight difference has an effect on the cooperative behaviour of the lions at these locations.

There are two primary elements in the relationship between lions and their prey that affect hunting behaviour – size and speed. Schaller, in discussing the speed relationship between a lion and its prey, notes that of all prey only the warthog has a lower speed than the lion, whilst the buffalo can achieve the same speed in escape as a lion. Schaller identifies the lion's top speed as 48 to 59 km/hr while Thomson's gazelle average 70 to 80 km/hr and the wildebeest 80 km/hr. Thus, when hunting the majority of prey, lions cannot rely on speed alone to achieve their goal. Interestingly, the buffalo does not need to use speed because a buffalo can cause a large amount of damage to a lion.

In general the prey's reaction to the presence of lions is not one of fear and panic. Schaller observed that the majority of the time when the lions were not hunting the prey was prepared to keep a reasonable distance from the lions and keep them under observation. The response to being attacked by a lion is for the herd to simply scatter, which can cause general confusion to lions. Schaller observed that sometimes the lions were unable to select an individual from the scattering herd resulting in the failure to achieve a kill. Lions use sight as the primary sense during hunting, although sound and smell may contribute to the initial location of prey in the Etosha national park where the biomass of potential prey is considerably lower. Schaller shows that group hunts that occur upwind from the prey have about 33% greater chance of success than those that occur down-wind from the prey.

Since lions are at a disadvantage when chasing most of their prey, they make use of the cover available to increase their chance of success. Schaller's observations of Serengeti lions shows that, although the majority of kills are in the open, prey are generally killed near areas that offer the greatest cover. The Etosha National Park is a flat arid plain with much less cover than the Serengeti plains, and yet even here the short grass is essential for the lions, for without some element of cover the lions have very little chance of gaining the advantage of surprise or the ability to ambush possible prey. The majority of hunts take place at night so that greater cover can be gained by use of the darkness.

2.2 The Predators

There are a number of factors that determine the use of group hunting in different prides of lions. For example, the environment that the lions live in has an effect. On the Serengeti plains, solitary hunts occur for approximately 48% of hunts, whilst in the Etosha National Park solitary hunting only occurred in 1% of hunts. This disparity is attributed to the difference in the environment and type of prey available in the two areas. The Etosha National Park is a vast semi-arid plain with little natural cover whilst the Serengeti National Park is a rich habitat. Faced with a reduction in cover those lions from Etosha seem to be forced into a situation were cooperative hunting provided a greater kill rate. The type of prey also has an effect on cooperative hunting. Over over 80% of the diet of the lions of the Etosha nation park is made up of large and/or fleet-footed prey. Serengeti lions have a greater variety of prey that includes animals, such as the warthog (the greatest occurrence of solitary hunting occurring in the Serengeti was upon the small warthog). In the Serengeti National Park, those prey which are hard to catch are hunted using group methods.

The available studies of the methods and tactics of group hunting adopted by lions give a similar basic plan of the hunting process. It starts when the group spots the prey, sometimes initiated by a single lion identifying the prey and looking at it, to which the other lions respond by looking in the same direction – the only clear form of "communication" evidenced in the hunting process. The group fans out, with certain lions stalking at a greater distance to encircle the prey. The encircling lions launch the *attack*, seemingly to drive the prey towards the other lions who *ambush* the prey from their cover position. A failed ambush may cause a *rush* after the prey for a short distance.

Stander's observations showed that, in general, lions followed approximately the same patterns when hunting. He

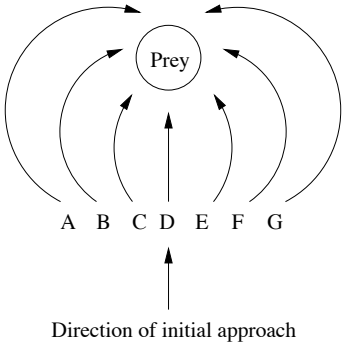


Figure 1: A schematic of generalised lion hunting behaviour.

divided the lions into seven different stalking roles, shown in Fig. 1, grouping these roles into Left Wing, Centre and Right Wing positions. Although no lions were exclusively fixed to any position and most hunts involve fewer lions than positions, Stander noted that the lions did seem to have a preference for certain positions, with 'wings' having a preferred angle of attack be it left or right. These positions would seem to be determined by the physical nature of the individual lions. The lions were measured and weighed, and it was discovered that the 'centres' were generally bigger, heavier and older than the 'wings'. In general the 'centre' lions are more capable of assuming different positions, but his research also showed that in situations when the lions were hunting in their preferred roles the success of the group increased by 9%. The hunting success rate also increased with the size of the group – the greater the size of the group the greater chance the lions had of assuming their preferred positions. Recent video evidence would suggest that the younger lions approach the prey faster than the older lions, possibly due to lack of experience of the younger lions or due to the younger lions needing to move further to get into position.

In general Stander observed that the average distance that the 'wings' stalk would be about 320m. Once the lions are in position, the breakdown of the roles in the hunt occurs with the 'wings' initiating the majority of the attacks. The 'centres' are more likely to be involved in 'ambushes' than the 'wings'. If the prey is not caught in an 'ambush' it would be the 'wings' that would become involved in the 'rush'. The 'rush' part of the hunt is only a short distance – Schaller observed that in general the lions could maintain a fast run for only a few hundreds meters before having to stop and pant.

3 Hypothesis

Although the details of the behaviour are known, the mechanism by which such behaviour is coordinated (if it is at all) is unknown. [Stander, 1991] suggests that the 'wings' could keep, as a 3rd point, the prey on an imaginary straight line between them, enabling the Lions to 'model' the locations of the other lions. Unfortunately, it is difficult to conceive that the Lions should need to locate lions on the other side of the encirclement in order to coordinate behaviour, since this would raise the possibility that the stalking lions could be spotted by the prey. However, Stander's proposals were the seeds of an alternative possible mechanism.

[Reynolds, 1987] used links of attraction and repulsion to produce the emergent group behaviour seen in flocking and herding, since applied to many other forms of emergent behaviour. We hypothesised that, through initial repulsion from the prey during stalking representing the need to remain unseen, attraction to the prey during stalking representing the need to get within range of the prey to attack, repulsion from other lions representing the need to obtain coverage of the prey's lines of escape, and attraction to other lions representing the need to close off possible escape between lions, an encircling behaviour during stalking would be seen. This hypothesised mechanism would require sight of the two neighbouring lions and the prey alone. Thus, no global knowledge and minimal local knowledge is required, and the behaviour would be an emergent property of the interaction of the agents involved.

The initial algorithm using this simple attraction-repulsion approach can be expressed simplistically as follows:

```
repeat until movement > chaseDistance or kill:
  if noticedMovement(sight) or lions.tracking < 2:
    lions = closestTwoLions
  if prey == null: prey = closestPrey
  if lions.tracking != 0 or prey != null:
    attractOrRepel(lions,prey)
  else: wander
  sight = vision(lions,prey)
```

4 Simulation

To investigate the proposed hypothesis a simulation was built in which distributed algorithms describing the action-selection and behavioural response of predators (the 'lions') and prey could be investigated. The intent of the simulation was to capture key aspects of the coordinated behaviour of the lions in the approach, stalk, rush and kill. Thus the simulation developed was intentionally simplistic, in marked contrast to the realism obtained from a complex simulation such as that of [Tu and Terzopoulos, 1994]. This decision is readily justified by recourse to a consideration of the nature and purpose of *simulation*; the oft-discussed consequences that such a decision implies are acknowledged.

In the simulation the following implementation decisions were made:

- The predators and prey are represented by a point-and-spread method, occupying a small circular area on a large 2-D torus simulation surface.
- Vision is the primary means of location of prey and other predators by the lions, and therefore must be modelled more accurately than simple line-of-sight. Vision is modelled as a cone extending forward from the location of the predator with a predefined spread. An animal in the cone is seen if $\lceil \frac{d}{d_{max}} g \rceil + (\frac{s_{max}}{2} - s + 1) + c < g$, where d is the distance to the target, g is the granularity of vision, s_{max} is the maximum speed of the animal, s is the current speed, and c is a current cover value, thereby trading-off visual acuity with distance, movement and cover. Peripheral vision is modelled as a second cone, overlaying the first, with less distance and a wider spread. Peripheral vision is sensitive to movement detection, but objects in peripheral vision will not

be 'recognised'. Prey have two main and peripheral vision cones directed 90 and 270 degrees from the direction the prey is facing. These are have less distance but are wider than those of the predator.

- Although scent is a factor in the prey locating predators, vision is the key component in identifying the close location of a predator and avoiding an attack. Similarly, scent is only a small factor in prey location by the predator and is not a primary factor in the final approach-stalk-rush behaviours. Therefore, scent is not modelled in the simulation.
- The predators are provided with four basic behaviours — *rest* - wandering as a part of a group in search of prey, *approach* - where the predator approaches a prey to observe it, *stalking* - where the predator uses cover and crouched movement to close the distance to the prey without detection, and *rush* - where the predator chases (for a short distance) and seeks to bring down the prey.
- A number of pre-determined triggers for behaviours are included that correspond to known single animal behaviour. The *approach* behaviour will be triggered when a predator spots a prey within attacking distance, or when a predator sees another predator approaching the prey. The *rush* behaviour will always be triggered when the prey flees away from a predator ([Schaller, 1972] identifies that lions launch a 'rush' opportunistically if they stumble upon a prey at close range or if the prey being stalked spots the predator and flees).
- The prey are provided with two basic behaviours — *grazing* - moving around a specific area looking for potential predators whilst eating, and *fleeing* - if a lion is spotted within a simulated distance of 50m [Schaller, 1972] or is spotted stalking or rushing, then the prey will turn away from the predator and flee. Fleeing is always away from the detected predators, although it is recognised that more complex fleeing strategies could be adopted [Cliff and Miller, 1996].
- It is assumed that the prey is always faster than the predators.
- In order to provide a measure of attraction / repulsion for use in the simulation the Lennard-Jones Potential was employed. Normally describing the relationship between water molecules, it provides a useful means of expressing the degree of attraction or repulsion as a function of distance between two points and is expressed as: $y = a * (\frac{b}{d})^{12} - (\frac{b}{d})^{6}$, where a is the rate of attraction/repulsion, b is the distance where the forces are in balance and d is distance between the animals. Although this could be used to vary the speed of approach to the balance point, we simply use it for the binary decision to approach or move away from an animal.

A simulation will start with predators placed spatially close to one another and within range for at least one of the predators to immediately detect the prey (i.e. there is no requirement within the simulation for the larger task of prey location). Each agent in the simulation (predator or prey) is independant, without any global coordination once the simulation starts and until the prey is killed or escapes. No global information is available to any agent in the simulation.

5 Investigation of the proposed algorithm

The algorithm identified in section 3 uses peripheral vision to identify when new movements are seen when the predator is looking at the prey or another predator. In testing it was found that since a predator has to turn to sight the prey or a predator, it can lose sight of a predator it was aware of and find another predator instead. This can cause the predator to spend its time searching for and trying to align with neighbouring predator (who are all moving) and seemingly lose interest in the prey.

A modification was made to the algorithm to prioritise stalking of the prey and to provide memory of the prey's last position so that the lion can re-locate the prey after it has turned. This version performed better, producing encircling behaviour at the desired attract/repel distance with a good distribution of predators around the prey. However these changes also introduced occasional problems. Where a predator B was directly between predator A and the prey, predator A would seek to move a back from predator B whilst being attracted towards the prey. Similarly predator B would seek to move towards the prey to get away from predator A whilst being repelled by the predator. This would result in a deadlock.

Although partially successful, it was also clear that the algorithm was not simulating some of the fundamental behaviour observed. The algorithm allowed predators no preference for position in the circle and did not normally show the younger predators covering larger distances. Rather than continue to modify the initial algorithm with additional constraints, an alternative approach seemed possible.

Lions within a pride display a form of dominance hierarchy, with older heavier lions typically dominant over younger lighter lions. In section 2.2 it was noted that [Stander, 1991] observed a preference for positions, with the 'wings' usually taken by the younger lions and the 'centre' taken by older and heavier lions. We hypothesised that this structure could be constructed using the same attraction / repulsion mechanisms if the mechanism included weighting for dominance (which might be argued to correlate with either a fear of more dominant lions, or a wish to avoid dispute over dominance) within a stalking area around the prey. We further identified that the faster movement of the wing lions was required to encourage rapid resolution of the dominance order without excessive 'shuffling' of positions.

Introducing this change in the algorithm not only demonstrated the desired encircling behaviour, but it was discovered that each predator only needs to be aware of the location of one immediate neighbour at a time in order to create the encircling behaviour. Since the maintenance of cover is important during stalking, this reduction in the requirements of sight suggests that the algorithm identified is plausible.

No information is available from the literature about how the final attack is initiated. This is an important matter, since early initiation will cause the prey to be chased into a location before the ambush is set. However, with the less dominant predators now being pushed to the wings and the more

dominant predators settling into position first it becomes easy to devise a trigger. The solution is in two parts. Firstly no predator can rush until the correct attract/repel distance with the prey and other predators is found, and secondly the distance away from the prey at which it is acceptable to initiate the 'rush' is weighted by dominance. Thus, less dominant predators will start the 'rush' as soon as they are in-place, at which time the other predators will also be in place. The older predators will not 'rush' until the prey is within a smaller distance, thereby creating the ambush. The final algorithm[2] was thus:

```
repeat until movement > chaseDistance or kill:
  prey, lion = closest(sight,memory)
  memory = [lion,prey]
  if distance(prey) in stalkBoundary and (fleeing(prey)
       or movement > stalkDistance(dominance)): rush
  elsif prey != null and lion != null:
    if collisionLikely(lion): moveAway(lion)
    elsif distance(prey) in stalkBoundary:
      attractOrRepelMax(lion,prey)
    else: attractOrRepel(lion,prey)
  elsif lion != null or prey != null:
    attractOrRepel(lion,prey)
  else: wander
  sight = vision(memory)
```

6 Results

Initial tests ran the final algorithm simulating the lions hunting behaviour ten times to identify whether the encirclement behaviour with appropriate distribution based on 'dominance' was consistently displayed. In these tests four lions were simulated, the average number involved in Stander's observations. In every test the predators were started in close proximity to one another and within sufficient distance to sight the the prey. Parameters were set as follows: $d_{max} = 900$ (lion), 140 (prey); $g = 12$; $c = 5$ (lion stalking), 1 (prey and lion not stalking); $s_{max} = 5$; $s = 2$ (lion wandering and stalking), 4 (lion rush), 1 (prey wander), 5 (prey fleeing), $a = 5$; $b = 100$ (prey) and 150 (lions). The encircling and rush behaviour is illustrated in figure 2, with the (randomly generated) dominance setting for the predator identified beside the corresponding plotted pathway.

The results in the plots show several different patterns depending on the success of the hunts. Successful hunts tended to look like those shown in plot (a) and (b). The correspondance between the 'dominance' of the simulated predators and the pathways shown on these plots illustrates the emergent order of the lions in the encircling behaviour. The initial rush behaviour by low dominance lions who drive the prey towards the more dominant lions, one of whom makes the kill, can also be seen in the plots.

A failed kill is illustrated by plot (c), which shows the situation where the prey notices the lions before any encirclement could be performed so that the prey flees from the lions. This example shows that a dominant lion does not give chase to the prey but the less dominant lions chase the prey for a distance. This behaviour emerges as a result of the lower distance to 'rush' for more dominant lions that otherwise provides the lion 'ambush' behaviour, and is similar to observed

[2]For a detailed discussion of the algorithm and its development, the reader is referred to [Dalrymple-Smith, 2003].

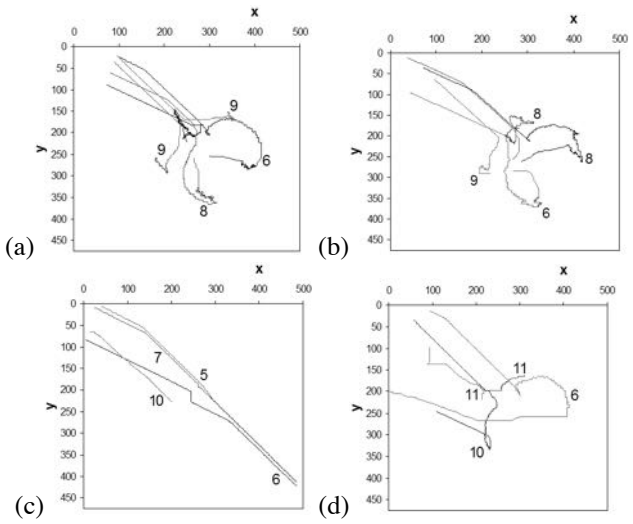


Figure 2: The encircling and rush of lions - (a) & (b) Successful hunts; (c) A hunt where the prey spotted a stalking predator; (d) A hunt where the prey escaped the ambush. Only the lion traces are shown, with the 'dominance' value indicated by each lion trace.

Table 1: Percentage of hunts with kills (a) for each dominance level, including the number of simulated hunts lions of that dominance were involved in, and (b) for each group size.

(a)

Dom	Hunts	Kill%
11	34	56
10	40	48
9	42	43
8	41	34
7	39	49
6	39	56
5	31	42

(b)

Lions	Kill%
6	53
5	60
4	27
3	40
Any	45

behaviour. Plot (d) also shows a failed kill. In this case the low dominance lion starts the chase and the prey flees, as expected. However all the remaining lions have a similar high dominance and therefore none will join the 'rush' until the prey is close. This allows the prey to run away from the single chasing lion, past the ambush and get away.

Another result (not shown on the plots) occurs on the occasions when the lions are out of position with the dominant lions on the 'wings' rather than the low dominance lions. In this situation the low dominance lions 'rush' from the central position enabling the prey to flee away from any ambush. This behaviour is interesting in relation to the observations discussed earlier that the probability of the lions having a successful hunt is dependant on the lions encircling at their preferred position.

A further investigation simulated 60 hunts with group sizes ranging from 3 to 6 predators (chosen to reflect the average group size of 4 lions, observed by Stander). The results are shown in table 1a and 1b.

Table 1a also shows that the percentage of successful hunts in relation to the hunts in which predators of a particular dom-

inance are involved. No clear pattern can be identified. It was hypothesised that this could be because the success or otherwise of a group hunt is dependant on the mix of dominance in the group. An analysis of the individual simulation results revealed that groups with more low dominance or more high dominance predators were less successful. This is readily explained by the observation that with a preponderance of low dominance predators it is unlikely that an ambush would be set up correctly, whilst with a preponderance of high dominance predators the prey is more likely to find a pathway of escape (see Figure 2d). Unfortunately no literature was available that would allow these findings to be compared with observed behaviour.

There is no clear pattern in the results shown in table 1b. It is interesting to note that [Stander, 1991] gives the success of hunts in the Etosha National Park at 27%, which is the same success rate seen in all simulations run with only four predators. Furthermore, Stander noted that the success of the hunt is increased with the group size, which is also seen with the overall kill rate in the simulation of 45%. However, in order to obtain statistically valid results a larger experimental run is necessary, and a fuller exploration of the parameter space of the simulator is required.

7 Conclusions and Further Work

The results presented in this paper represent the output of preliminary investigations of a highly simplified simulation to test the hypothesis that the group predatory behaviour of lions can be mimicked using attraction / repulsion between the lions and their prey. Although a full investigation, particularly of the parameterisation of the simulation, has yet to be conducted, the results demonstrate that many of the observed behaviours in the group hunting of lions are replicated in the simulation.

It would be inappropriate to claim that the proposed mechanism is therefore the mechanism deployed by lions in group hunting. However, it is important to note that the complex cooperative behaviour demonstrated here emerges from the deployment of very simple dynamic interactions. It is of particular interest that the introduction of dominance relations was not only key to producing behaviour that correctly mimicked observed stalking, rush and ambush behaviour, but was also key to an important simplification in the visual communication required for agent coordination. This finding opens an avenue of further research for multi-agent coordination.

We hypothesise that such 'cooperative' behaviour within a group hunting situation could arise from the 'selfish' behaviour of lions, and be deployed in a coordinated manner, when use is made of group dominance relations. Verification of this hypothesis from further live observations would help to resolve the debate on the nature of the 'cooperation' seen in lion hunting, and would provide further insights into the role of dominance in animal groups.

Further work is required to investigate the parameterisation of the model used and then to use the simulator to obtain a wider range of results to which statistical analysis can be applied. Future work will seek to explore the role of dominance in multi-agent coordination in distributed computer applications, and the identification of dominance agreement algorithms for fault tolerant systems.

References

[Boden, 1996] M. A. Boden. *The Philosophy of Artificial Life*. Oxford University Press, Oxford, UK, 1996.

[Cao *et al*., 1994] Y. Uny Cao, Alex S. Fukunaga, Andrew B. Kahng, and Frank Meng. Cooperative Mobile Robotics: Antecedents and Directions. In *UNKNOWN*, 1994.

[Cliff and Miller, 1996] D. Cliff and G. F. Miller. Co-Evolution of Pursuit and Evasion II: Simulation Methods and Results. In P. Maes, M. Mataric, J-A. Meyer, J. Pollack, and S. W. Wilson, editors, *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior (SAB96)*, pages 506–515. MIT Press Bradford Books, 1996.

[Creel, 1996] S. Creel. Communal hunting and pack size in African wild dogs Lycaon pictus. *Animal Behavior*, 50:1325–1339, 1996.

[Creel, 1997] S. Creel. Group hunting and group size: assumptions and currencies. *Animal Behavior*, 54:1319–1324, 1997.

[Dalrymple-Smith, 2003] Hugo Dalrymple-Smith. Simulating Lions Hunting Behaviour. Technical report, The Department of Computer Science, The University of Bath, UK, 2003.

[Griffin, 1984] D. R. Griffin. *Animal Thinking*. Harvard University Press, Cambridge, Mass., 1984.

[McFarland, 1994] D. McFarland. Towards Robot Cooperation. In *Proceedings of International Conference on the Simulation of Adaptive Behaviour*, 1994.

[Reynolds, 1987] C. W. Reynolds. Flocks, Herds and Schools: A distributed behavioural model. *Computer Graphics*, 21(4):25–34, 1987.

[Schaller, 1972] G. B. Schaller. *The Serengetti Lion*. University of Chicago Press, Chicago, 1972.

[Scheel and Packer, 1991] D. Scheel and C. Packer. Group hunting behaviour of lions: a search for cooperation. *Animal Behavior*, 41:697–709, 1991.

[Stander, 1991] P. E. Stander. Cooperative hunting in lions: the role of the individual. *Behav. Ecol. Sociobiol.*, 29:445–454, 1991.

[Tu and Terzopooulos, 1994] Xiaoyuan Tu and Demetri Terzopooulos. Artificial Fishes: Physics, Locomotion, Perception, Behavior. In *Proceedings of ACM SIGGRAPH'94*, pages 43–50, Orlando, Florida, 1994. ACM Computer Graphics Proceedings.

[Zaera *et al*., 1996] Nahum Zaera, Dave Cliff, and Janet Bruten. (Not) Evolving Collective Behaviours in Synthetic Fish. In *Proceedings of International Conference on the Simulation of Adaptive Behavior*, pages 635–644. MIT Press, 1996.

Having it both ways – the impact of fear on eating and fleeing in virtual flocking animals

Carlos Delgado Mata
Bonaterra University, Mexico

Ruth Aylett
Heriot-Watt University, UK

Abstract

The paper investigates the role of an affective system as part of an ethologically-inspired action-selection mechanism for virtual animals in a 3D interactive graphics environment. It discusses the integration of emotion with flocking and grazing behaviour and a mechanism for communicating emotion between animals; develops a metric for analyzing the collective behaviour of the animals and its complexity and shows that emotion reduces the complexity of behaviour and thus mediates between individual and collective behaviour.

1 Introduction

For much of its history, Artificial Intelligence (AI) had stressed reasoning and logic and almost ignored the role of emotion in intelligent behaviour. Minsky (1985) was one of the first to emphasise the importance of emotion for Artificial Intelligence. Since then, affective systems for embodied autonomous agents, robotic and graphical, have become an expanding research area. Approaches divide roughly into two: low-level accounts, focusing on animals in general, sub-symbolic behavioural architectures and neuro-physiologically inspired [Velasquez 1997, Canamero 1998], and high-level accounts, focusing on humans, symbolic appraisal-driven architectures, and inspired by cognitive science [Ortony et al 1988, Gratch et al 2001]. In this work, we concentrate on a low-level account, applied to exemplary flocking mammals (sheep, deer), and demonstrate the role of fear as a social regulator between individual and group behaviour. We take the set of 'primitive emotions' namely: anger, fear, disgust, surprise, happiness and sadness, [Ekman 1982] as a plausible set for other mammals than humans, and examine how they can be integrated into an ethologically-based action-selection mechanism.

An evolutionary approach to emotions suggests that for affective systems to have developed and remained under the pressure of selection, they must play a definite functional role within the overall architecture of animals. A number of such functions can be identified in relation to the selection of actions. One is to modify behaviour: a sheep that experiences an anxiety-inducing stimulus may carry on grazing but bunch up more tightly with the rest of the flock. A second is to switch behaviours: a sheep experiencing a threatening stimulus inside its flight zone will flee. A third is to avoid dithering between competing behaviours by adding weight to one of them [Blumberg 1994], and a fourth and related function is to sustain a selected behaviour for an appropriate interval – a fleeing animal can typically no longer perceive the threatening predator, but fear keeps it running, in effect acting like a cheap short-term memory.

However flocking animals do not behave merely as individuals, they engage in the collective behaviour known as flocking. Reynolds [1987] showed that flocking behaviour does not require a complex internal architecture but can be produced by a small set of simple rules. In his model of so-called *boids*, every individual (boid) tries to fulfil three conditions: cohesion or flock centring (attempt to stay close to nearby flockmates), alignment or velocity matching (attempt to match velocity with nearby flockmates), and separation or collision avoidance (avoid collisions with nearby flockmates). Flocking is thus a collective emergent behaviour.

This approach has produced sufficiently believable collective behaviour to be used for stampedes in a number of animated films. Nevertheless, mammals do in fact have a complex internal architecture, unlike social insects, and a wide range of individual behaviours: a motivation for this work was to reconcile the generation of collective behaviour by a small set of rules with the more complex agent architecture required for a mammalian behaviour repertoire.

An important behaviour in the ungulate repertoire is grazing, requiring spatial orientation behaviours. Two such mechanisms of particular relevance are described in Lorenz [1981]. The first, *kinesis* can be summarized as a reactive rule of slowing down when encountering favourable conditions and speeding up for unfavourable ones: this can also be related to escape behaviour. However most organisms do not move in an absolutely straight line; when orienting to favourable localities: the effect of kinesis can be improved by increasing the angle of the random deviations from the straight line, and these are inherent to locomotion in any case. By these means, the organism is kept in the desirable environment longer and is made to exploit an increased part of its area, especially relevant to grazing. This second enhanced mechanism is termed *klinokinesis* and it is found in grazing mammals, as well as in swimming protozoa and higher crustacea. This represents an important example of

individually-oriented behaviour which conflicts with the rule-set for flocking.

2 An ethologically inspired action-selection mechanism

The work discussed here has been implemented with graphically-embodied flocking animals (sheep, deer) in a 3D interactive virtual environment. In order to test the hypothesis that an affective system can act as a regulating mechanism between individual and social behaviour, an ethologically-motivated architecture was developed for the virtual animals.

The basic task of a virtual animal brain has often been split into the three sub-tasks of perception (sensing the environment and interpreting the sensory signals to provide a high-level description of the environment), action selection (using the perceptual and emotional inputs to decide which of the animal's repertoire of actions is most suitable at that moment) and motor control (transforming the chosen action into a pattern of "physical" actions to produce the animation of the animal). To this we add a fourth subtask: generating emotions (affecting the behaviour of the animals, exemplified by the conspecifics flight-flocking), Figure 1 shows a detailed diagram of the designed architecture developed as a result, and the next sections describe its components.

While not claiming neurophysiological accuracy, the architecture splits its overall functionality across biologically-plausible subsystems. Thus the module *hypothalamus* is used to store the drives (for example, hunger), the *sensorial cortex* stores sensor data, the *amygdala* contains the emotional systems such as Fear, Joy and Anger, and *Basal Ganglia* contains the hierarchical mechanism for selecting actions, similar to those described by ethologists. Each of the listed modules is defined in XML giving the name of each of the system/variables, the inputs associated to them, a weight, and a function (acting as a filter, in most cases a sigmoid) which in turn generated a feed-forward hierarchy like the one described by Tyrrell [1993]

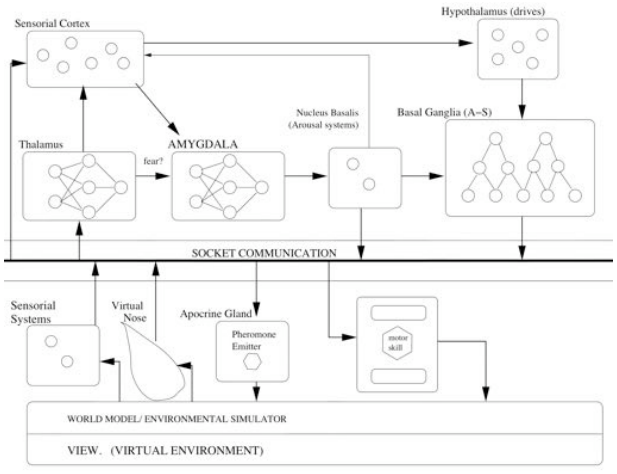


FIGURE 1: Complete architecture

2.1 Communicating emotion

Taking the position that emotion partly functions as a communication mechanism, a novel feature of this work is that the perceptual component has been designed to support the communication of emotion among conspecifics. In the real world, emotional transmission is almost certainly multimodal, with certain modes such as the perception of motion being particularly difficult to model. Thus we have limited ourselves for now to a single mode, and the one we have chosen is pheromones, perceived by a virtual olfaction sensor.

Recent experiments [Grammer 1993] have shown that mammals, including humans, emit pheromones through apocrine glands as an emotional response, and as means to communicate that state to conspecifics, who can adapt their behaviour accordingly; research has found that odours produce a range of emotion responses in animals, including humans [Izard1993]. This is adaptively advantageous because olfaction is part of the old smell-brain which can generate fast emotional responses, that is without the need of cognitive processes.

Grammer [1993] argues that every living creature has a distinctive molecular signature that can be carried in the wind, variously showing it to be nutritious, poisonous, sexual partner, predator or prey. Neary [2001] points out that sheep, particularly range sheep, will usually move more readily into the wind than with the wind, allowing them to utilise their sense of smell.

Our architecture models the exteroceptors used by real animals to detect the presence of chemicals in the external environment as a virtual nose. An environmental simulator has been developed: its tasks include changing the temperature and other environmental variables depending on the time of day and on the season, using statistical historical data. An alarmed animal sends virtual pheromones to the environmental simulator and they are simulated using the free expansion gas formula in which the volume depends on the temperature and altitude (both simulated environmental variables). The expansion of the pheromone cloud at timestep=9 can be seen in a graphical environment in Figure 6 below. To compute the distribution of the pheromones a set of particles has been simulated using the Boltzmann distribution formula:

$$n(y) = n_o e^{\frac{mgy}{k_b T}}$$

Here m is the pheromone mass; g is the gravity; y is the altitude; k_b is the Boltzmann number; T is the temperature; n_0 is N/V where N is the number of molecules exuded from the apocrine gland (related to the intensity of the emotion) and V is the volume. The virtual nose detects pheromones from a threshold of 200.10^{-16} reflecting values taken from the relevant literature.

2.2 Action-selection mechanism

The problem of action selection is that of choosing at each moment in time the most appropriate action out of a repertoire of possible actions. The process of making this deci-

sion takes into account many stimuli, including in this case the animal's emotional state.

Action selection algorithms have been proposed by both ethologists and computer scientists. Models suggested by the former are usually at a conceptual level, while those of the latter (with some exceptions – see Tyrrell:1993, Blumberg:1994) generally do not take into account classical ethological theories. Dawkins [1976] suggests that a hierarchical structure represents an essential organising principle of complex behaviours: a view shared by many ethologists [Baerends:1976, Tinbergen:1969].

Recent research has found that the Basal Ganglia plays an important role in mammalian action selection [Montes-Gonzalez 2001] and our mechanism is implemented in the *Basal Ganglia* module in Figure 1 as a three-level tree. To avoid sensory congestion, each of Top, Intermediate and Bottom nodes receives sensor data directly as well as data from a higher-level node. Actions are selected by Bottom nodes, dispatching them via a UDP socket to the Animation engine located in the *Body* module of Figure 1. Figure 2 shows the overall interconnections in the animal brain.

TABLE 1: Finite State Acceptor for klinokinesis

State	Input	Resulting state
start	go-default	stand-still
stand-still	P(0.3)	walking
stand-still	P(0.3)	starting-to-eat
stand-still	P(0.2)	rotating-left
stand-still	P(0.2)	rotating-right
stand-still	in-fear	end
stand-still	do-nothing	stand-still
walking	P(0.3)	stand-still
walking	P(0.7)	walking
rotating-left	P(0.9)	stand-still
rotating-left	P(0.1)	rotating-left
rotating-right	P(0.9)	stand-still
rotating-right	P(0.1)	rotating-right
starting-to-eat	head-down	eating
eating	P(0.6)	eating
eating	P(0.4)	finishing-eating
finishing-eating	head-up	stand-still

This mechanism is based on Tyrrell [1993] who in turn developed Rosenblatt and Payton's original idea [1989] of a connectionist, hierarchical, feed-forward network, to which temporal and uncertainty penalties were added, and for which a more specific rule for combination of preferences was produced. Note that among other stimuli, our action selection mechanism takes the emotional states (outputs of the emotional devices) of the virtual animal.

Klinokinesis was modelled as a Finite State Acceptor [Arkin1999], augmented with transitions based on probability, as seen in Table 1.

2.3 The flocking mechanism

The basic Reynolds rules of cohesion, alignment and separation have been extended with an additional rule (escape) in which the virtual animal moves away from potential danger (essentially, predators) in its vicinity. More importantly, the flocking behaviour itself is parameterised by the emotional devices output, that is, by the values of the emotions the virtual animals feel. Therefore, in our model each virtual animal moves itself along a vector, which is the resultant of four component vectors, one for each of the behavioural rules.

The calculation of the resultant vector, V(elocity), for a virtual animal A is as follows:

$$V_A=(Cf \cdot C_ef \cdot Cv)+(Af \cdot A_ef \cdot Av)+(Sf \cdot S_ef \cdot Sv) +(Ef \cdot E_ef \cdot Ev) \quad (2)$$

$$Velocity_A=limit(V_A, (MV_ef \cdot MaxVelocity)) \quad (3)$$

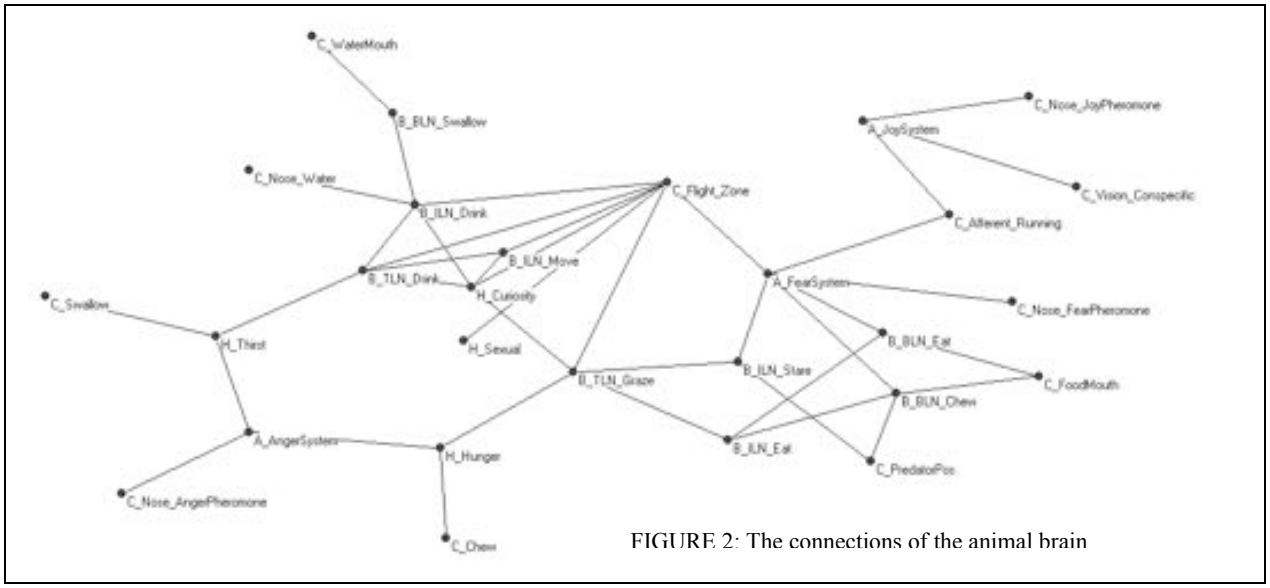


FIGURE 2: The connections of the animal brain

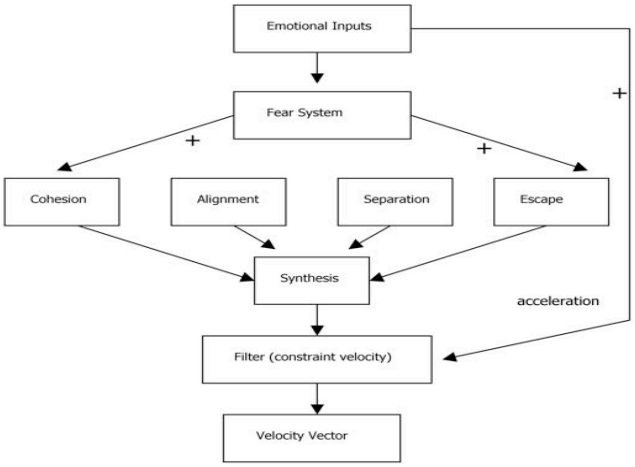


FIGURE 3: Influence of Fear on Flocking

where Cv, Av, Sv and Ev are the component vectors corresponding to the cohesion, alignment, separation and escape rules respectively. Cf, Af, Sf and Ef are factors representing the importance of the component vectors Cv, Av, Sv and Ev and allow weighting of each component vector independently. In our current implementation they can be varied, in real time, from a user interface. $C_e f$, $A_e f$, $S_e f$ and $E_e f$ are factors representing the importance of the respective component vectors given the current emotional state of the virtual animal. Each of these factors is a function that takes the current values of the animal's emotions and generates a weight for its related component vector. MaxVelocity is the maximum velocity allowed to the animal. In the current implementation it can be varied from a user interface. $MV_e f$ is a factor whose value is calculated as a function of the current values of the animal's emotions. It allows the increase and decrease the animal's MaxVelocity depending on its emotional state as shown in Figure 3. limit is a function whose value is equal to the greater of its two parameters.

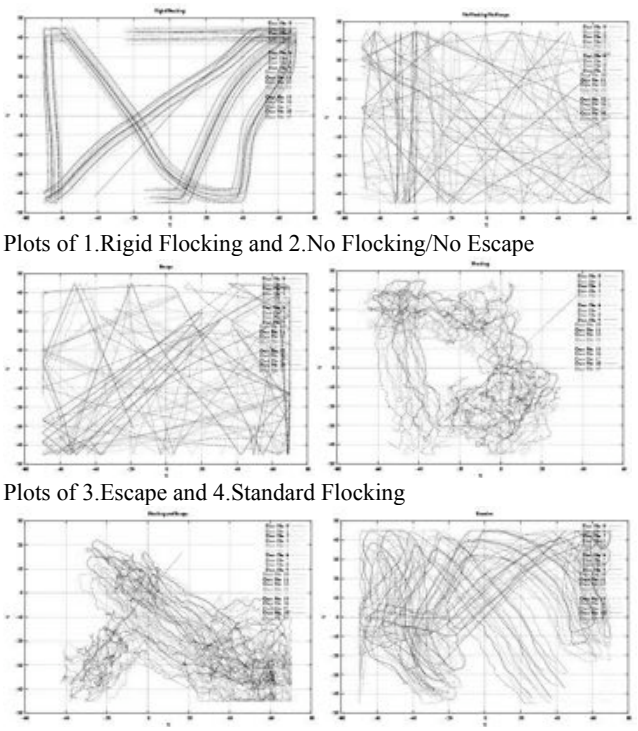
The emotional factors ($C_e f$, $A_e f$, $S_e f$, $E_e f$, and $MV_e f$) reflect ethological heuristic rules. For example, the greater the fear an animal feels, the greater the weight of both its cohesion vector (it tries to stay closer to nearby flockmates) and its escape vector (it tries to stay farther from the potential danger). The resultant vector obtained by adding the four basic vectors is then scaled so as not to exceed the maximum speed. Note that maximum velocity is also parameterised by fear: the greater the fear an animal feels, the greater the speed it is able to reach.

3. Evaluating the emergent behaviour

3.1 Experimental data

Our hypothesis that fear can serve as a regulator between individual and social behaviour was evaluated through an experiment in which 5,10, 15 and 20 animals were plotted over 600 timesteps for the following six conditions:

1. *Rigid Flocking*. The herd of animals was tightly packed (maximum10 centimetres distance between each) and animals were all facing the same direction at all times. This is the baseline condition for optimum coordination.
2. *No Flocking No Escape*. Each animal moved on its own with no knowledge (perception) of other animals or predators. This is the baseline condition for individual behaviour.
3. *Escape*. Similar to the previous scenario except that animals perceive predators and individually move to avoid them.
4. *Standard flocking*. Animals perceive each other, try to avoid collisions between each other and try to stay close to the herd.
5. *Standard flocking with Escape.* As the previous case but animals perceive predators, and move to avoid them.
6. *Escape with emotion*. Emotion (fear) is elicited and communicated amongst animals via artificial pheromones when predators are perceived.



Plots of 1.Rigid Flocking and 2.No Flocking/No Escape

Plots of 3.Escape and 4.Standard Flocking

Plots of 5.Standard Flocking/Escape and 6.Emotion/Escape

FIGURE 4: Plots of 6 cases, 600 timesteps, 20 animals

Figure 4 shows the trajectories plotted for the 20 animals case, and it is intuitively clear to the eye even at this very low resolution that very different patterns of behaviour are being produced. What is required is a way of assessing the complexity of the emergent behaviour in each case.

3.2 An evaluation mechanism

We follow the approach of Wright et al. [2001] who presented a method for characterising the pattern of emergent behaviour and its complexity using singular values and entropy.

In the matrix A below, M = 600 (number of samples) and N = 4 (degrees of freedom: position x,y and velocity x,y):

$$A = \begin{pmatrix} {x_1}^1 & {y_1}^1 & \dot{x}_1^1 & \dot{y}_1^1 & \cdots & {x_1}^N & {y_1}^N & \dot{x}_1^N & \dot{y}_1^N \\ & & & & \vdots & & & & \\ {x_M}^1 & {y_M}^1 & \dot{x}_M^1 & \dot{y}_M^1 & \cdots & {x_M}^N & {y_M}^N & \dot{x}_M^N & \dot{y}_M^N \end{pmatrix}$$

To compute the singular values, the following equation from linear algebra is used:

$$A = USV^T$$

The singular values $\sigma_i = S_i$ are all non-negative and generally are presented in a decreasing sequence $\sigma_1 >= \sigma_2 >= \ldots >= \sigma_N >= 0$; singular values can be used as an approximation of the matrix. We do not have space to display the singular values for 5,10,15,20 animals for all six cases here, but if they are represented in bar chart form they show that each flocking case has its own distinctive shape.

The next step is to compute the entropy from the N singular values which are normalised, because by definition $\Sigma_i P_i = 1$ [Bonabeau et al.1999]: in our case P_i is σ_i. The following equation is used to calculate entropy:

$$E_s = -\sum_{i=1}^{N} {\sigma_i}' \log_2 {\sigma_i}'$$

where σ'_i. is the normalised singular value. And since entropy can be seen as a $\log_2$ count of the number of states in a system [Bonabeau et al.1999], the effective number of states and thus the complexity is given by the expression:

$$\Omega = 2^{E_s}$$

3.3 Results

Figure 5 above shows a plot of the complexities for different types of flocking with different number of animals. It can be seen that rigid flocking (bottom line) shows the least complexity, intuitively supported by looking at Figure 4, top left. Flocking; flocking with escape; no flocking, no escape; and escape behaviours (top four lines) are more complex

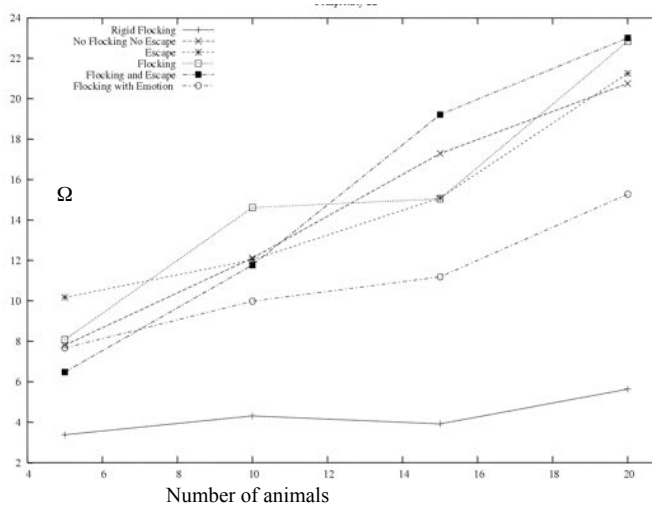


FIGURE 5: Plot of complexity (Ω) against animal numbers for 6 cases

FIGURE 6: Pheromone cloud visualised at timestep = 9

than rigid flocking, but they are also almost always more complex than flocking with emotion (second line up).

The exception is the 5-animal case where flocking with emotion, is more complex than flocking with escape. This can be explained by a further set of experiments carried out in which it is shown that at least nine animals are needed to maintain flocking behaviour. With fewer than this, when the animals escape from a predator, some separate from the flock and do not regroup at all during the 600 time-steps.

Thus we conclude that the introduction of an emotional system into action-selection, where emotion can be transmitted between animals, is to mediate between the complexity of individual behaviour and the rigidity of collective behaviour. It allows a dynamic trade-off between spreading widely, advantageous in seeking new grass to graze - inherent in klinokinesis - and staying together, advantageous in the case of attack by predators. Emotion in this case acts as a social regulator for flocking animals, demonstrating that it has an important social function in addition to its already-understood role in regulating individual behaviour.

In addition to the 2D-tracking of trajectory just described, the virtual animals have also been implemented in a real-time 3D graphical world, which can be run in a 4-sided immersive display system (or CAVE). The implementation consists of nearly 28,000 lines of C++ code of which 10,949 implement the brain. Figure 6 shows a screen-shot of a sheep in a graphical world illustrating the spread of the pheromone cloud at timestep = 9. A further objective of the work discussed elsewhere [Delgado-Mata et al 2003] is to examine how far the presence of emotionally-driven autonomous animals can increase the feeling of immersion experienced by a human user in such environments.

4. Conclusions and further work

We have presented an ethologically-inspired virtual animal architecture in which primitive emotions have been incorporated into action-selection and a method for communicating emotion between animals using virtual pheromones has been included, allowing the extension of the classic approach to flocking to incorporate emotion. We have shown

that the effect of adding the emotional input to flocking together with the communication mechanism is to reduce the complexity of individual behaviour without requiring rigid lock-stepping. This substantiates the hypothesis that emotion mediates social behaviour, underlining the functional role of affect in action-selection.

Extensions to this work might include individual variation in animals, both across characteristics like fearfulness, and across gender: there is evidence that ewes spend more time grazing and rams significantly longer lying. The presence of lambs would also introduce an interesting element of social heterogeneity, while animals with other behavioural responses to predators – musk ox for example form an outward facing ring – could be explored.

The use of 3D space in this implementation is limited to the pheromone propagation algorithm: both perception and locomotion were implemented as 2D mechanisms. Given mammals have significantly less mobility in 3D than the classic examples of fish or birds, a more realistic application of manoeuvrability constraints would not only look more natural but might also have practical implementations for flock fragmentation in the face of predators. A classic predator strategy is to peel off an individual flock member, and including one or more intelligent predators would allow predator-prey interaction to be investigated.

Finally, although the architecture developed targetted animals such as sheep and deer rather than humans, the extension of the approach into emotionally-driven human crowds would open up a much larger field of investigation.

Acknowledgements

We would like to acknowledge the help of Rocio Ruiz Rodarte of ITESM, Mexico City, in allowing us to use the beautiful virtual world of Palenque and also ConaCyt for the financial support of Carlos Delgado Mata.

References

[Arkin1999] Ronald C Arkin. Behavior-Based Robotics. MIT Press, Cambridge,Massachusetts, 1999

[Baerends 1976] G. P. Baerends, "The functional organization of behaviour," *Animal Behaviour*, no. 24, pp. 726–738, 1976.

[Blumberg 1994] Bruce Blumberg, Action-selection in hamsterdam: Lessons from ethology, Proc 3rd Int Conf. on Simulation of Adaptive Behavior SAB 94, 1994

[Bonabeau et al.1999] Eric Bonabeau, Marco Dorigo, and Guy Theraulaz. Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press, USA, 1999.

[Canamero 1998] Canamero, D. Modeling Motivations and Emotions as a Basis for Intelligent Behaviour, Proc1st Int. Conf. on Autonomous Agents. ACM Press pp148—155, 1998

[Dawkins 1976] Richard Dawkins, Hierarchical organisation: A candidate principle for ethology, Growing Points in Ethology (Bateson & Hinde, ed.), Cambridge University Press, 1976.

[Delgado-Mata et al 2003] Delgado-Mata, C; Ibanez, J and Aylett, R.S. Let's run for it: conspecific emotional flocking triggered by virtual pheromones. Proc. Smart Graphics 2003, LNCS 2733, Springer-Verlag 2003

[Ekman 1982] Ekman, P. Emotions on the Human Face. Cambridge University Press

[Gratch et al 2001] Gratch, J; Rickel, J; & Marsalla, S. Tears and Fears, *5Th Int. Conf. on Autonomous Agents*, ACM Press, pp113-118 2001

[Grammer 1993] Karl Grammer, 5-alpha-androst-16en-3alpha-one: a male pheromone ?, *Ethology and Sociobiology* 14 (1993), no. 3, 201–207.

[Ibanez et al, 2004] Ibanez, J, Delgado, C, Aylett, R.S. & Ruiz, R. (2004) Don't you escape. I'll tell you my story. *Proceedings, MICAI 2004,* Mexico City, April 2004 Springer Verlag LNAI 2972 pp49-58

[Izard, 1993] Carol~E Izard, Four systems for emotion activation: Cognitive and noncognitive processes, *Psychological Review* (1993), no.1, 68--90.

[Minsky 1985] Marvin Minsky, The society of mind, Simon & Schuster, New York, USA, 1985

[Montes-Gonzalez 2001] Fernando M Montes-Gonzalez. An action Selection Mechanism based on Vertebrate Basal Ganglia. PhD thesis, Psychology Department, University of Sheffield, Sheffield, United Kingdom, 2001.

[Neary, 2001] M.Neary, Sheep sense, *The Working Border Collie* www.working-border-collie.com/article3.html

[Ortony et al 1988] Ortony, A; Clore, G. L. and Collins, A. The Cognitive Structure of Emotions, Cambridge University Press 1988

[Reynolds 1987] Craig W.Reynolds, Flocks, herds, and schools: A distributed behavioral model, *Computer Graphics* 21 (1987), no. 4, 25–34

[Rosenblatt and Payton 1989] J.K. Rosenblatt and D.W. Payton, A fine-grained alternative to the subsumption architecture for mobile robot control, *Proc. IEEE/INNS IJCNN* (Washington DC), v 2, June 1989, pp. 317– 324.

[Tinbergen 1969] Nikolaas Tinbergen, The study of instinct, Oxford University Press, United Kingdom, 1969.

[Tyrell 1993] Toby Tyrrell, Computational mechanisms for action selection, Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland, 1993

[Valasquez 1997] Juan Valasquez Modeling emotions and other motivations in synthetic agents. *Proc 14th Nat Conf on AI, AAAI 1997,* MIT/AAAI Press 1997

[Wright et al.2001] W A Wright, R E Smith, M Danek, and P Greenway. A generalisable measure of self-organisation and emergence. In G. Dorner, H. Bischof, and K. Hornik (eds) *Artificial Neural Networks – ICANN 2001*, LNCS 2130, pp 857-864. Springer-Verlag, 2001

Building agents to understand infant attachment behaviour

Dean Petters
University of Birmingham
School of Computer Science
Birmingham B15 2TT, United Kingdom
d.d.petters@cs.bham.ac.uk

Abstract

This paper reports on an autonomous agent simulation of infant attachment behaviour. The behaviours simulated have been observed in home environments and in a controlled laboratory procedure called the Strange Situation Experiment. The Avoidant, Secure and Ambivalent styles of behaviour seen in these studies are outlined, and then abstracted to their core elements to act as a specification of requirements for the simulation. A reactive agent architecture demonstrates that these patterns of behaviour can be learnt from reinforcement signals without recourse to deliberative mechanisms.

1 Introduction

This paper describes work which aims to further understanding of infant attachment behaviour using an AI 'design-based' approach. This means explaining the structures that would be required in a system's design to enable attachment phenomena to be produced. It follows from Petters (2004), which explains how a generic 'design-based' approach was adapted to the specific purpose of simulating attachment phenomena. The particular attachment behaviour under investigation is the pattern of infant response to separations from and subsequent reunions with their carers in a controlled procedure that occurs in an unfamiliar laboratory environment. This procedure is known as the 'Strange Situation Experiment'. The simulation will be attempting to explain: the different patterns of infant behaviour found in separation and reunion episodes of the Strange Situation; why patterns of reunion behaviour match patterns of home behaviour most closely; and why the Strange Situation and home behaviour, when considered across all subjects, cluster to give three main categories of attachment style in infancy. The elements needed for a specification of requirements are abstracted from empirically observed behaviours, and can be presented as a set of mini-scenarios which the autonomous agent simulations need to fulfil. Since autonomous agent techniques are based upon simplified models of complete systems that endure over time, infant and carer agents in the simulation can respond to each other's behaviours in a dynamic and adaptive manner.

An architecture has been implemented, which when placed in appropriate training environments, can simulate the formation of the three attachment styles being investigated. This architecture is entirely reactive, which means it possesses no mechanisms that allow it to 'look-ahead' and predict the effect of its actions. It is implemented at a high, goal-oriented level. In addition to the objective of explaining attachment, this project aims to act as a 'test-bed', and compare the performance of a variety of architectures. These will differ with regard to the possession of a number of capabilities, such as whether the architectures can perform evaluations of possible actions or whether they have access to, or can use explicit forms of representation.

Internal validity is achieved for each candidate architecture by ensuring that it fulfils the requirements set out by the scenario. Any architecture that can reproduce the behaviour required by the scenario has passed a form of sufficiency test and is a 'proof of concept' for that theory (Cooper, 2002). External validity is related to how well the specification of requirements, in the form of a scenario, represents the behaviour we are trying to explain. Architectures that fully reproduce the scenario can be assessed against each other, and may differ in how they fulfil the scenario, whether by principled or ad-hoc means. Additional constraints can be derived from a wealth of linked empirical data and theory from cross-species, evolutionary, neurosphysiological and cross-cultural branches of Attachment Theory.

2 The nature of the problem: behaviours to be explained

The Strange Situation Experiment is not strictly an experiment but rather it is a standardised laboratory procedure that presents infants with a controlled and replicable set of experiences. What the infant experiences over the eight short episodes of this procedure is intended to activate and intensify infant attachment behaviours, which might include the infant looking more at the carer, moving towards the carer or crying to gain the carer's proximity. The effect of the Strange Situation procedure is designed to be similar to situations that infants commonly encounter in real life, with the important qualifications that each infant taking part should experiences the same environment and that the infants responses are recorded by video through a one way mirror.

The Strange Situation procedure was originally devised to investigate differences in attachment behaviour observed in a comparison of infants from rural villages in Uganda and infants from suburban communities in Baltimore, USA. When under observation, the Ugandan infants exhibited significantly elevated levels of anxious behaviour, despite being in familiar surroundings. Ainsworth *et al*. (1978) hypothesised that the intense behaviours seen in the Ugandan study might be evoked more incisively if the Baltimore infants were put in an unfamiliar environment. The Strange Situation was therefore created, and in the first study it was carried out on one year old infants who had previously undergone extensive observation in the uncontrolled and familiar environment of their homes. Importantly, the presence of extensive observations at home meant that individual difference classifications could be made not only based upon infant responses in the 24 minutes duration of the entire Strange Situation procedure, but also in response to the 72 hours of home observations made in the preceding year. The close matching of behaviour in the Strange Situation and the home studies has allowed the Strange Situation procedure to act as an indicator of the quality of the mother-child relationship that exists outside of the laboratory (Goldberg, 2000).

In total, in the the home and Strange Situation observation stages of the study, 23 different types of infant behaviour and 26 types of maternal behaviour were observed. The reported aspects of infant behaviour included data on the prevalence of specific behaviours such as: frequency and duration of crying; differing responses to mother's comings and goings; behaviour relevant to contact; and anger with the mother. In the Strange Situation additional infant data was gained on the nature of the infant's exploration and responses to the stranger. Detailed data on the mother was recorded only for behaviour at home and included data on specific measures of behaviour and codings that captured more abstract patterns of behaviour across superficially different types of interaction. The specific measures of behaviour included: how many times and for how long mothers did not respond to the infant crying; the duration and affectionate quality of pick-up episodes; the mother's level of aversion to physical contact; and the frequency of unpleasant physical contact being provided. Abstracted patterns of maternal behaviour were coded from across different types of episode and included the level of emotional expression and general sensitivity. The mother's behaviour in the Strange Situation was controlled by the testers and so was similar in all cases.

The eight episodes of the Strange Situation, which are all of three minute duration, are: (1) mother and infant introduced to unfamiliar room; (2) mother is nonparticipant while infant explores; (3) a stranger enters; (4) mother leaves but stranger remains, (the first separation episode); (5) mother returns and stranger leaves, (the first reunion episode); (6) mother leaves infant on its own, (second separation episode); (7) stranger returns; and (8) mother returns and stranger leaves, (second reunion episode). When the results were evaluated using a Multiple Discriminant Function Analysis the infant responses were found to be clustered into three major categories of attachment style, labelled: Avoidant (type A), Secure (type B) and Ambivalent/resistant (type C) (Ainsworth *et al*., 1978).

A meta-analysis of cross-cultural patterns for 2000 Strange Situation classifications across 8 countries found that the original Baltimore study, and most other studies (from countries including the US, China, West Germany, Great Britain, Netherlands, Sweden and Japan) fitted into a group where about two thirds of infants were assigned to the Secure (B) category, a fifth assigned to the Avoidant (A) category and an eighth to the Ambivalent (C) category (van Ijzendoorn and Kroonenberg, 1988). Studies with statistically outlying distribution patterns included: Israeli and US studies with elevated proportions of Ambivalent infants; a West German study with higher numbers of Avoidant infants; and a Japanese study where the number of Avoidant infants was lower than the international average and the number of Ambivalent infants higher. Since the original studies, a fourth attachment category has been formed, labelled the Disorganised/disoriented (D) style. These infants are found in small numbers in non-clinical samples, and usually come from home environments with particularly inadequate care (Goldberg, 2000). At this preliminary stage of the project the cross-cultural and type D disorganised infant data will be set aside and the simulations will concentrate on providing explanatory models of how the other three attachment types come to be formed. However, these studies are a rich source of constraints for future evaluation of the architecture described here.

What we want to do is understand why styles of attachment behaviour form as clusters, rather than being distributed more evenly along some spectrum of behaviour. We also want to understand the architectural mechanisms by which these behaviours come about, and to form a theory explaining the purpose, if any, of these behaviours for the infant. We cannot use every detail of all the behaviours recorded in any one study, as this would result in overfitting of the data and poor generalisation. Also, trying to implement observations regarding such things as a mother's sensitivity to the infant's food preferences would result in arbitrary details of no relevance to our central theoretical questions. We need to be selective and find a level of abstraction, that captures the essence of the empirical results, and that can form a scenario that will then act as a specification of requirements for the purposes of evaluation. The separation behaviours are not a clear guide in this regard. This is because separation behaviours in the Strange Situation laboratory setting are not fully predicted by the carer's and infant's behaviour in the home environment. Importantly, however, reunion behaviour in the Strange Situation is strongly predicted by the home behaviour of the mother and the infant (Meins, 1997).

Regardless of how they reacted in separation;

- the infants whose response to their mothers on reunion in the Strange Situation was: to not seek contact or avoid their mother's gaze or physical contact with her, *are described as insecure-avoidant and labelled type A*. These children return quickly to play and exploration but do so with less concentration than secure children. Whilst playing they stay close to and keep an eye on their carer. They received care at home which can be summarised as being consistently less sensitive. In comparison with average levels across all groups: A type

mothers were observed at home being less emotionally expressive and having a greater aversion to close physical contact; they left infants crying for longer durations and provided more physical contact of an unpleasant nature; and at home these infants were more angry, they cried more and were observed to 'sink-in' less during physical contact.

- the infants whose response to their mothers on reunion was: positive, greeting, approaching, making or accepting contact with, or being comforted by her, *are described as securely attached and labelled type B*. These children returned to play and exploration in the room quickly. They received care at home which can be summarised as being consistently more sensitive. In comparison with average levels across all groups: B type mothers were observed at home being more emotionally expressive and having a smaller aversion to close physical contact; they left infants crying for shorter durations and provided less physical contact of an unpleasant nature; at home these infants were less angry, they cried less and were observed to 'sink-in' more during physical contact.
- the infants whose response to their mothers on reunion was: not being comforted and overly passive or showing anger towards their mothers, *are described as insecure-resistant/ambivalent and labelled type C*. These children do not return quickly to exploration and play. They received care at home which can be summarised as being less sensitive and particularly inconsistent. In comparison with average levels across all groups: C type mothers were observed at home being more emotionally expressive and having a smaller aversion to close physical contact; they provided physical contact which was unpleasant at a level intermediate between A and B carers and left infants crying for longer durations; at home these infant's were more angry, they cried more but were observed to 'sink-in' more during physical contact.
- the infants whose response to their mothers on reunion was: totally disorganised and confused, *are described as insecure-disorganised and labelled type D*. The home environment of behaviour for this very small proportion of infants has been found to be dysfunctional, often with depressed mothers or with maltreatment of the infant (Meins (1997), Ainsworth *et al*. (1978), and Weinfield *et al*. (1999)).

From empirical observation to scenario.

From these results we need to focus on those behaviours of particular importance for the purposes of creating a scenario that will act as a specification of requirements. Together Avoidant and Ambivalent infants are termed Insecure infants. Although infants of these types differ in a number of respects, they have much more in common with each other than either has in common with the Secure infants. The Secure versus Insecure distinction is therefore the clearest distinction made in this study. The second distinction concerns why some infants end up classified as Insecure Avoidant whereas others are classified as Insecure Ambivalent.

The differences between Secure and Insecure infant-carer pairs that we want to include in mini-scenarios are:

- In reunion episodes of the Strange Situation, Secure infants show some distress but get back to play quicker and with more attention in their play.
- At home, Secure infants communicate with less intense negative tone, show less angry protest, crying less frequently and for shorter durations, and are more rewarded by close physical contract.
- At home, the carers of Secure type infants respond in a timely fashion to more communications from the infant.

The differences between Avoidant and Ambivalent infant-carer pairs that we want to include in mini-scenarios are:

- In reunion episodes of the Strange Situation, Avoidant infants show little angry protest but Ambivalent infants show particularly resistant and angry behaviour.
- Both Avoidant and Ambivalent infants show more angry protest at home, they both cry more frequently and for longer durations. What separates their home behaviour is that Ambivalent infants are more rewarded by close physical contact than are Avoidant infants.
- At home, the carers of Avoidant infants reliably reject infant signals indicative of a desire for closeness and when they do make physical contact it is more often unpleasant. The carers of Ambivalent infants also frequently reject signals indicative of a desire for closeness, but this pattern of behaviour is more inconsistent. These carers vary in the quality of physical contact they provide.

Putting aside theories based upon innate temperament.

This section will briefly set out the evidence that theories of innate temperament make as a claim to partially explain Strange Situation behaviour. However, these theories will then put aside apart from as consideration for future work. Strong evidence for innate temperamental traits as major causal factors for the infant behaviour in the Strange Situation emerged after the first study by Ainsworth *et al*. (1978). This first study found a strong correlation between maternal behaviour at home and infant behaviour in the Strange Situation, and a weaker correlation between early infant behaviour and infant behaviour in the Strange Situation. However, later studies have failed to confirm the clear difference in the magnitudes of these correlations. In a meta-analysis of thirteen different studies Goldsmith and Alansky (1987) found that although the correlation between home maternal behaviour and infant Strange Situation behaviour was stronger than that between infant home behaviour and infant Strange Situation behaviour, the gap had narrowed from that found in Ainsworth *et al*.'s study.

It has been suggested that the avoidant infants don't show distress in reunion episodes simply because they are not distressed, and this lack of distress is the result of innate temperamental differences (Goldberg, 2000). Studies that measured the physiological correlates of stress for infants undergoing the Strange Situation have been carried out (Hertsgaard *et al*., 1995; Spangler and Grossman, 1993). Using

heart rate and cortisol measurements as indicators of covert stress these studies found that the stress levels of avoidant infants were at least as high as the secure and ambivalent groups. However, it might still be argued that avoidant infants show less overt distress than secure and ambivalent infants because of an innate difference in their ability to manage stress. In addition, from cross-species studies of attachment, Suomi (1999) reports that innate differences similar to human temperamental traits give rise to three categories of attachment behaviour in Rhesus Monkeys. These categories appear to fulfil some of the same roles and functions as that of the Avoidant/Secure/Ambivalent distinction found in humans, and unlike with human infants these categories are innately determined and invariant through-out ontogenetic development.

The simulation must ultimately have the power to represent the full range of likely phylogenetic and ontogenetic causal factors. For now this work will focus on models that concentrate on caregiving influences on infant behaviour. These effects seem to be the largest and models of innate temperament may also be better incorporated when the model is deepened to include lower level implementational details.

3 An initial solution: a reactive architecture

The aim of this project is to build a number of infant and carer software agents that reproduce the differing patterns of home and laboratory behaviour found in studies of attachment. The simulated agents should be designed in a manner that increases our understanding of how and why the patterns are formed in reality. The simulation needs to be as simple as possible, whilst being powerful enough to represent plausible competing mechanisms and possible causal structures that underlie the patterns of behaviour. The simulations reproduce the attachment behaviours in question at an abstract level, and do not replicate the lower level details of sensory modalities and motor actions.

Each simulation is split into a training period, corresponding to the lengthy home observations, and a test period, corresponding to the much shorter Strange Situation assessment. A mapping between input data (in the form of the carer's behaviour at home) to output data (in the form of the infant's behaviour in the Strange Situation stage of the simulation) emerges from the dynamic interaction of infant and carer agents in a 2-dimensional virtual world. In the training period infants explore the space of behaviours open to them, attempting to optimise their behavioural 'policy' in terms of plausible adaptive benefits, and these learnt 'policies' are carried forward to the Strange Situation stage of the simulation.

A reactive architecture has been implemented and reproduces the different attachment behaviours in the scenario. The design of this system is inspired by structures, mechanisms and functions described in Ethology and Attachment Theory, particularly the Behavioural System architecture described by Bowlby (1969 1982) and the analysis of avoidant and ambivalent strategies outlined, respectively, by Main and Weston (1982) and Cassidy and Berlin (1994).

The architecture has three major divisions: perceptual subsystems, a central selection and arbitration subsystem, and

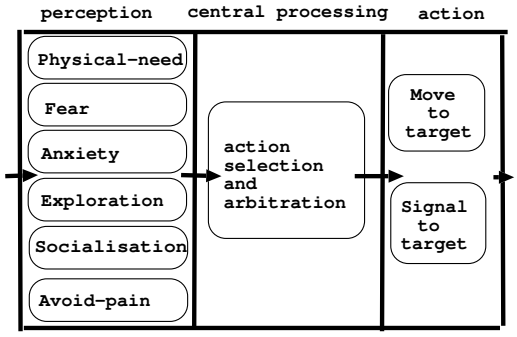


Figure 1: A Reactive Design. This architectural hypothesis postulates that; at one year of age, all Strange Situation patterns of behaviour are produced without resort to deliberative mechanisms. This hypothesis does not preclude simple deliberative processes occurring in other behavioural domains.

action subsystems. (see figure 1). There are six perceptual subsystems, which each have an implicit goal, and each providing proposals for action with a variable activation level. There are also two action subsystems which are not mutually exclusive, and can be active simultaneously. In between the perceptual subsystems and the action subsystems is the selection and arbitration mechanism, which selects the group of actions with the highest activation that do not exclude each other being carried out.

The six perceptual systems have the implicit goals of: maintaining physical requirements (an abstraction of food, warmth and cleanliness); maintaining safety from unfamiliar objects; maintaining safety from remoteness of the carer; learning about objects; learning about agents; and avoiding harm from previously unpleasant objects, agents or events. These subsystems will henceforth be termed, respectively, the Physical-need, Fear, Anxiety, Exploration, Socialisation and Avoid-pain perceptual systems, these terms being labels, not definitions of properties. The two action subsystems are Moving and Signalling, each of which has a target. When the targets of the actions are included there are nine different atomic actions possible, six movement actions and three signalling actions: to move to the carer (Move-carer); to signal to the carer (Signal-carer); move to the stranger (Move-stranger); signal to the stranger (Signal-stranger); to move to a target toy object (Move-toy); to not signal (Not-signal); and to move away from the carer, stranger or toy (Away-carer, Away-stranger and Away-toy). Figure 2. lists the perceptual subsystems alongside the actions that each may activate.

If the purpose of the simulation was to model attachment behaviour in infancy without regard to individual difference data then the architectural mechanisms described above would be sufficient. The simulation provides mechanisms for how infants may adapt to their carer's behaviour. Before we describe the learning mechanisms present in the infant we will summarise the three types of carer behaviour in the simulation.

Carers in the simulation have a small repertoire of actions, and different patterns of caregiving are distinguished by the

Physical-need	**Move-carer** **Signal-carer**
Fear	**Move-carer** **Signal-carer**
Anxiety	**Move-carer** **Signal-carer**
Exploration	**Move-toy** **Not-signal**
Socialisation	**Move-carer** **Signal-carer** **Move-stranger** **Signal-stranger**
Avoid-pain	**Not-signal** **Away-carer** **Away-stranger** **Away-toy**

Figure 2: Shows the mapping from the six 'Behaviour system' perceptual modules to the actions that might possible be activated by those modules.

timings, thresholds and effectiveness with which these actions are carried out. Carers are either concerned with foraging for energy or with attending to the infants in some manner. Carers possess three key parameters that guide their behaviour. These parameters are labelled: Feed-self-now, Respond-to-infant and Panic-now. Carers use up energy as they move around. If a carer has an energy level below the Feed-self-now threshold then it will always choose foraging instead of any signalling to, or energy provision for, the infant. This threshold is low in the case of the carers of Secure and Ambivalent infants, so that these carers rarely interrupt or delay their caregiving to forage or harvest energy. This is not the case with Avoidant carers, who possess a higher value of the Feed-self-now parameter, and who therefore often interrupt attending to the infants, whatever the infants level of distress, to forage for energy to build up their own energy levels. The available energy in the environment can be varied. If energy levels are abundant then the carers of Avoidant infants behave more like the other carers. However if energy is scarce the reverse happens, and all carers seem like carers of Avoidant infants. This accords with an empirical study that found low levels of support to caregivers predicted a higher likelihood of infants receiving an Insecure classification (Crockenberg, 1981).

Infants can emit signals with a range of affective tone, representing both smiles and distress. The Respond-to-infant parameter is set so that the carers of Secure and Ambivalent infants will ordinarily be very prompt in their responses to infant signals. The carers of Avoidant infants respond to signals at a higher threshold than the other carers. An important exception to these patterns of behaviour is caused by the actions of the Panic-now parameter. This parameter sets the level at which the carers' behaviour switches to a different and less efficient mode. The carers of Secure and Avoidant infants have a very low probability of 'panic' being triggered, so that the other three parameters provide a good summary of their behaviour. The carers of Ambivalent carers often 'panic', and this means that their behaviour is patchy, unpredictable and overall much less responsive than that provided by the carers of Secure infants. The notion of 'panic' is not a good representation of the ordinary use of this term, but does reflect the qualities of the carers of Ambivalent infants found in studies that have used the Adult Attachment Interview (AAI) to assess the attachment status of infant caregivers (Hesse, 1999). The AAI analyses the discourse properties of adult carers talking about attachment relations and describes the carers of Secure infants as autonomous and 'free to evaluate' and the carers of Avoidant infants as dismissing. The carers of Avoidant infants are described as preoccupied, enmeshed, and often engaged with angry struggles with their own carers. It is a pattern of caring which results from these states of mind that the 'panic' behaviour is trying to capture.

In reunion episodes, why do secure infants return to attentive exploration sooner?

Secure infants get back to attentive exploration earlier than other infants because their Anxiety subsystem is less activated. In an unfamiliar environment the Exploration subsystem will always posses at least moderate activation and when the activation of the Anxiety subsystem drops below this level the behaviour switches to exploration. Another way of saying this is that, although they have just undergone a distressing separation, Secure infants feel safer in the reunion episodes than the other infants do. They feel free to explore because they assess they are under less threat. This is because Secure infants have learnt in their previous home experiences that their carers are reliable providers of security.

All the infants assess security in their Anxiety subsystems by reference to a parameter called the Safe-range distance. When the distance to their carer is less than the Safe-range the Anxiety system passes no activation and the infant can be said to be feeling no insecurity. When the carer travels beyond this threshold, so that the distance from carer to infant is greater than the Safe-range distance, the Anxiety subsystem starts to pass an increasingly high level of activation. The longer the carer stays beyond the Safe-range limit the higher the activation level goes. Eventually the Anxiety subsystem gains control of behaviour and issues Move-carer and Signal-carer actions that should bring the carer closer.

In the training stage of the simulation there are repeated instances where the carer goes beyond the Safe-range limit, is called back, and then responds promptly or otherwise. The Safe-range limit is then updated by a re-inforcement signal ($\mathcal{R}(t)$), that is a function of the time (t) the carer takes to respond from the infant's first signalling, and is given by equation 1:

$$\mathcal{R}(t) = \frac{\alpha}{1 + e^{\beta(t-(t^m+t^k))}} - \gamma t \qquad (1)$$

The constants α and γ control the maximum rewards and punishments, respectively, that the infant receives. The constant β sets the gradient of the decrease of the reward, a small positive value for β produces a gradual decrease in the size of reward each time step, higher positive values for β produce decreases in reward that approach a step function from maximum to minimum reward over small periods of time.

The time step where the steepest decline in reward occurs is set by the terms t^m and t^k. The term t^m is the minimum possible time that a carer could respond, its inclusion means that infants do not expect carers to respond faster than the laws of the virtual world allow. The term t^k is a constant. When the time elapsed from bid to reward is less than $(t^m + t^k)$ the positive reinforcement is large. Another way of saying this is, prompt responses give large reinforcement signals and this results in the Safe-range distance being increased. When responses take more time reinforcement signals become negative 'punishment' signals. In this context 'punishment' doesn't mean that the actions that were taken are less likely to be taken in future. Quite the opposite occurs. The Safe-range limit is reduced by the value of the punishment signal. Therefore distances that are previously considered by the infant to be safe, are now beyond the Safe-range distance. The carer still has to forage and may still need to go as far afield in the future, so the chances are that after a decrease in Safe-range the carer will be less responsive in future.

If a number of decreases in the Safe-range distance occurred without any intervening prompt responses, the infant may become chronically untrusting of the carer's performance. The reverse obviously holds true for carers that carry out a series of prompt returns. This positive feedback mechanism, operating over a long training period, may be what drives the infant-carer pairs into the Secure/Insecure clustering seen in the Strange Situation studies. A carer whose performance is initially intermediate between Secure and Insecure may come to be perceived as at either extreme of caregiving. Figure 3 illustrates the results of computational experiments to find if there exists a level of carer responsiveness which is intermediate between secure and insecure forms of caregiving. On first inspection it may seem that a carer response threshold of 29 gives an intermediate Safe-range value. But what actually occurs at this value is a bifurcation between some experiments that show secure patterns and some that show insecure patterns, with a resulting intermediate average with high standard deviation. This has implications for the kind of predictions we might expect this model to produce. Figure 4 shows the results of ten experiments, where identical carers, all with a response threshold of 29, were matched with ten identical infants. In these experiments, five infants ended up secure and five ended up insecure, the difference due to small random variations that occured early in each simulated run.

What affect does possesion of a small Safe-range value have on an infant's perception of its carer's behaviour, and therefore its own consequent actions? Figure 5 shows how contrasting infant behaviours result from situations where carers are currently making identical relative movements. In the secure scenario, the carer is moving towards food but is still within the infant's Safe-range so the infant is not signalling and is moving towards an unexplored toy object. In the insecure scenario, the infant's past experiences have given rise to a smaller Safe-range. The carer has already crossed this boundary and the infant has switched goals from exploration to bringing the carer back within its Safe-range by signalling and moving towards the carer.

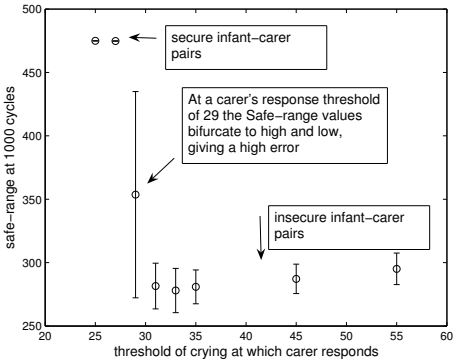


Figure 3: A snapshot of the infant Safe-range found at 1000 cycles for different values of carer responsiveness. Each data point is a mean from 20 simulations.

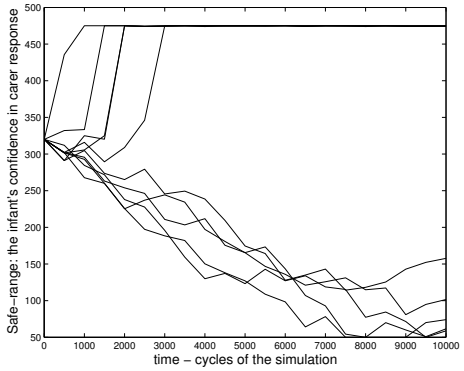


Figure 4: Infant Safe-range changing over time as carer responsiveness set at 29. Each experiment had an identical initial state, differences resulting from random elements early in the 10000 cycles.

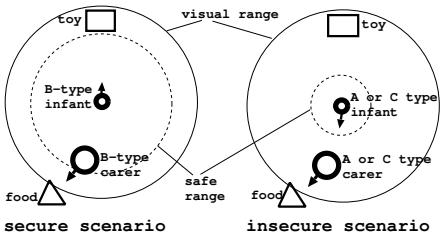


Figure 5: A secure infant moves towards a toy as its carer moves away towards food. When an insecure infant experiences the same event it moves and signals towards the carer.

Sroufe and Waters (1977) introduced the term 'felt-security' to emphasise infants are not merely measuring a one dimensional distance between infant and carer. Other factors, such as carer attentiveness, are used by the infant to measure security. We can view the Safe-range parameter as an abstract representation of a much richer measure of security than exists in reality. This does not change the core of our theory, whereby incidences of high or low carer responsiveness change the criteria by which the level of future security is assessed.

At home, why do Avoidant and Ambivalent infants both get angry and cry more?

Individual episodes of anger are initiated in the training phase of the simulation whenever three conditions are met. These are that the infant experiences an undesirable event, this event violates its current expectation, and the event has been brought about by the infant's carer. The key type of event that repeatedly gives rise to anger is when the infant signals for the carer's attention and proximity but the carer does not respond within the expected time-frame. Anger is not initiated immediately, but commences after the expected time for response has elapsed. The expected time for response is represented as the sum of the t^m and t^k terms in equation 1. Therefore the frequency of angry experiences is related to the learning of the Safe-range distance. The infant experiences anger in each separation episode where the time the carer takes in responding (t in equation 1) is greater than the sum $t^m + t^k$.

In reunion episodes, why do Avoidant infants show less anger and crying?

Instead of protesting when reunited with their carer, Avoidant infants return to exploration, but do so with less attention than Secure infants. Behaviour of this type has been described as a 'displacement activity'(Main and Weston, 1982). An example of displacement activity from animal behaviour might be found when an animal is faced with a con-specific with which it might fight with, or flee from, but instead starts to groom itself (Bowlby, 1969 1982). Displacement activities occur when two strongly activated behaviours 'cancel each other out', and a seemingly inappropriate behaviour becomes active. According to Ainsworth *et al*. (1978), Avoidant infants act avoidantly in reunion episodes to avert close physical contact.

The ability to avert close physical contact of an unpleasant nature has been implemented in the simulation by the inclusion of the Avoid-pain subsystem, which learns how rewarding close physical contact has been for the infant. Secure and Ambivalent infants receive rewarding experiences when in close physical contact with their carers. Avoidant infants receive less pleasant physical contact and are less rewarded. However, Avoidant infants still seek contact when their Physical-need and Anxiety activation levels are very high. The perceptual systems have bounds to their activations, and these are set so that the Avoid-pain subsystem does not override Physical-need or Anxiety subsystems when these subsystems are highly activated. This explains why the behaviours in the reunion episodes are indicators of the quality of the infant-carer relationship beyond the laboratory. Reunion episodes provide Anxiety activations that are just low enough that the balancing effect of the Avoid-pain subsystem can be seen. Insecure Avoidant infants differ in their reunion behaviour from Insecure Ambivalent infants because at close distances to their carers the goal of avoiding close contact inhibits active expression of secure goal behaviour, leading to behaviours linked to the exploration goal being activated as displacement behaviours.

The Avoid-pain subsystem 'solves' the problem of Avoidant reunion behaviour but creates a new problem of how to represent physical contact in this high-level simulation. A more satisfactory solution for future work would be for the distinctions in behaviour to emerge from the functioning of a theoretically grounded perceptual system, such as that reviewed by Polan and Hofer (1999). Rat studies have uncovered mechanisms related to physical contact and attachment, that are termed 'hidden regulators', and which are believed to be the physiological basis of the state of 'felt-security'. These mechanisms are believed to produce low level expectations regarding comfort and safety in rats and human infants.

4 Conclusion

This work uses autonomous agent techniques to create a psychological model of social interaction in infancy. To the author's knowledge it is the first software implementation that explains the three principal types of infant attachment as adaptations to caregiving style. The simulation captures fine grained and long term temporal properties of behaviour at an abstract level of description, and it shows how different attachment styles are formed and can become self-sustaining. The architecture might be implemented at a less abstract level of description by incorporation of the mechanisms of action selection found in the Basal Ganglia and described by Gurney *et al*. (2001) and reinforcement learning mechanisms described by Schultz *et al*. (1997). This simulation differs from computational models of infant development based upon stand-alone neural networks and production systems because it models a whole system which includes perception, action and internal processes embedded in a dynamic environment (Shultz, 2003). It also differs from many developmentally oriented agent-based simulations and other relevant work, such as social interaction in infant-like robots, because of its concentration on central processing and impoverished perceptual capacity (Schlesinger, 2001; Breazeal and Scassellati, 2000; Likhachev and Arkin, 2000).

Comparison with a model of adult Contention Scheduling is instructive: the infant architecture can be viewed as more abstract, more simple but also a broader version of the architecture described by Cooper and Shallice (2000). Both link perception to action, but the architecture for Contention Scheduling does so via a hierarchically organised network of action schemas, which represent goals and multiple levels of subgoals. The architecture for infant attachment is flat, with one level of goals activating a single level of atomic actions. These actions are also more abstract than those found in the model of Contention Scheduling. Future work may involve augmenting the two infant actions of moving and signalling

with the inclusion of less abstract and more numerous actions which may need to be respresented at multiple levels of abstraction, and hence require hierarchical organisation. There is a sense that the model of Contention Scheduling is aimed at a narrower set of phenomena than the ultimate objective for the infant architecture. Whereas Contention Scheduling is intended to only deal with routine activities, the infant architecture is intended to model responses to novel situations. To more fully realise this aim a hybrid architecture is in the process of creation, and this architecture is intended to integrate deliberative capabilities with the reactive action selection mechanisms described above.

The main contributions that this work makes are that it describes and then abstracts a set of data observed from attachment studies of human infants. A model of action selection based upon a reactive architecture has been implemented and evaluated against behavioural and physiological data. Further systematic exploration of its design space will be carried out. However, in analysing the simulation it is important to distinguish between its theoretical and its implementational assumptions. For example, many of the parameters described in the infant and carer architectures, such as the carer response threshold, do not have a straightforward translation to phenomena in reality. This work aims to contribute to multiple disciplines and demonstrates that the domain of attachment behaviour, in humans and other species, provides a valuable 'test-bed' for comparing the performance of different action selection mechanisms.

Acknowledgment

Thanks to Aaron Sloman, for much help and encouragement.

References

M. Ainsworth, M. Blehar, E. Waters, and S. Wall. *Patterns of Attachment: a psychological study of the strange situation*. Erlbaum, Hillsdale, NJ, 1978.

J. Bowlby. *Attachment and loss: volume 1 attachment*. Basic books, New York, 1969-1982.

C. Breazeal and B. Scassellati. Infant-like social interactions between a robot and a human caretaker. *Adaptive Behavior*, 8:49–74, 2000.

J. Cassidy and L.J. Berlin. The insecure/ambivalent pattern of attachment: Theory and research. *Child Development*, 65(4):971–91, 1994.

R. Cooper and T. Shallice. Contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, 17(4):297–338, 2000.

R. Cooper. *Modelling High-Level Cognitive Processes*. Lawrence Erlbaum Associates, New Jersey, 2002.

S.B. Crockenberg. Infant irritability, mother responsiveness, and social support influences on the security of infant-mother attachment. *Child Development*, 52:857–69, 1981.

S. Goldberg. *Attachment and Development*. Arnold, London, 2000.

H.H. Goldsmith and J. Alansky. Maternal and infant temperamental predictors of attachment: a meta-analytic review. *Journal of clinical and consulting psychology*, 55:805–16, 1987.

K.N. Gurney, T.J. Prescott, and P. Redgrave. A computational model of action selection in the basal ganglia II. Analysis and simulation of behaviour . *Biological Cybernetics*, 84:411–423, 2001.

L. Hertsgaard, M. Gunnar, M. F. Erickson, and M. Nachmias. Adrenocortical responses to the strange situation in infants with disorganised/disoriented attachment relationships. *Child Development*, 66:1100–1106, 1995. 4.

E. Hesse. The adult attachment interview, historical and current perspectives. In Handbook of Attachment, *eds. J. Cassidy & P.R. Shaver*, pages 395–433. Guilford Press, London, 1999.

M. Likhachev and R.C. Arkin. Robotic comfort zones. In *Proceedings of SPIE: Sensor Fusion and Decentralized Control in Robotic Systems*, pages 27–41, 2000.

M. Main and D.R. Weston. Avoidance of the attachment figure in infancy. In The place of attachment in human behavior, *eds. M. Parkes & J. Stevenson-Hinde*, pages 31–59. Basic Books, New York, 1982.

E. Meins. *Security of attachment and the social development of cognition*. Psychology Press, Hove, 1997.

D. Petters. Simulating infant-carer relationship dynamics. In *Proc AAAI Spring Symposium 2004: Architectures for Modeling Emotion - Cross-Disciplinary Foundations*, number SS-04-02 in AAAI Technical reports, pages 114–122, Menlo Park, CA, 2004.

H. J. Polan and M.A. Hofer. Psychobiological origins of infant attachment and separation responses. In Handbook of Attachment, *eds. J. Cassidy & P.R. Shaver*, pages 162–180. Guilford Press, London, 1999.

M. Schlesinger. The agent-based approach: A new direction for computational models of development. *Developmental Review*, 21:121–146, 2001.

W. Schultz, P. Dayan, and P.R. Montague. A neural substrate of prediction and reward. *Science*, 275:1593–1599, 1997.

T. Shultz. *Computational Developmental Psychology*. Bradford books, London, UK, 2003.

G. Spangler and K.E. Grossman. Biobehavioural organisation in securely and insecurely attached infants. *Child Development*, 64:1439–1450, 1993. 5.

L.A. Sroufe and E. Waters. Attachment as an organisational construct. *Child Development*, 48(4):1184–99, 1977.

S. J. Suomi. Attachment in rhesus monkeys. In Handbook of Attachment, *eds. J. Cassidy & P.R. Shaver*, pages 181–197. Guilford Press, London, 1999.

M.H. van Ijzendoorn and P.M. Kroonenberg. Cross-cultural patterns of attachment: A meta-analysis of the strange situation. *Child Development*, 59(1):147–156, 1988.

N.S. Weinfield, L.A. Sroufe, B. Egeland, and E.A. Carlson. The nature of individual differences in infant-caregiver attachment. In Handbook of Attachment, *eds. J. Cassidy & P.R. Shaver*, pages 68–88. Guilford Press, London, 1999.

Simulation, Emotion and Information Processing: Computational Investigations of the Regulative Role of Pleasure in Adaptive Behavior

Joost Broekens and Fons J. Verbeek
University of Leiden
Leiden Institute of Advanced Computer Science,
Leiden, The Netherlands.
{broekens, fverbeek}@liacs.nl

Abstract

Emotion plays an important role in thinking. In this paper we focus on the regulatory influence of pleasure on information processing in simulated adaptive agents. Our agent's pleasure is a function of its performance on the tasks it executes in the environment. Our model is based on *Reinforcement Learning* and the *Simulation Hypothesis*. The main hypothesis tested is: *if action-selection-bias is induced by an amount of simulated anticipatory behavior, and if this amount is dynamically controlled by pleasure feedback, then this provides additional survival value to an agent compared to a static amount of simulation*. Experimental results illustrate that this hypothesis holds true. Dynamic adaptation results in a learning performance that at least equals static simulation strategies, and it results in a major decrease of mental effort required for this performance. This is relevant to the evolutionary plausibility of the simulation hypothesis, for increased adaptation at lower cost is an evolutionary advantageous feature. In addition, our results provide clues of a relation between the simulation hypothesis and emotion.

1 Introduction

Emotion plays an important role in thinking. Evidence ranging from philosophy [Griffith, 1999] through cognitive psychology [Frijda, *et al.*, 2000] to cognitive neuroscience [Damasio, 1994; Davidson, 2000] and behavioral neuroscience [Berridge, 2003; Rolls, 2000] shows that emotion—in whatever form—is both constructive and destructive to a wide variety of cognitive phenomena. Normal emotional functioning seems to be necessary for normal cognition.

In this research we focus on the low-level influence of emotion on information processing in simulated adaptive agents. We define emotion as a combination of pleasure and arousal factors [Russell, 2003]. The agent's arousal is based on a metadescription of its memory, e.g., prediction accuracy. Pleasure is a function of the agent's relative performance on the tasks it executes in the environment. The agent uses Reinforcement Learning (RL) [Sutton and Barto, 1996]. In this paper we focus on the influence of pleasure as feedback to control the amount of simulated anticipatory behavior the agent uses to bias action selection. This influence is measured in terms of learning performance and total effort spent on simulated and overt interaction. Thus, we investigate the influence on learning if emotion is used to control the cognitive mechanism (i.e., simulation) that biases action-selection. We do not model categories of emotions nor use such emotions as information in symbolic-like reasoning. Reasons for our low-level approach include:

First, because emotion is integrated at multiple levels of processing and higher—conscious, reflective reasoning—levels have not always existed throughout evolution, one would expect an evolutionary advantage to integration at levels close to reward systems and behavioral control. On higher levels, emotion *regulates* information processing. Could emotion play such role at lower levels?

Second, from a computational point of view lower levels tend be more generic. Therefore, regulative mechanisms found can be applied to a wider area of disciplines including cognitive science and machine learning, for example meta-learning—how to autonomously monitor and, if necessary, adapt the learning mechanism used by the agent in order to better cope with the current task. If emotion is considered as a meta-learning system [Doya, 2000], it can be used to enhance artificial adaptive agents in a generic way. Regulative mechanisms that operate on higher cognitive levels may need a more complex concept of emotion or a dedicated cognitive architecture, and are therefore less generic.

Third, a low-level interpretation allows us to stay close to behavioral control and action-selection mechanisms thereby avoiding philosophical debates about emotion. Consequently, we use a modest—but broadly usable and less controversial—concept of emotion as basis for the research.

Fourth, Montague *et al*. [2004] recently argued that computational models of RL can be used to model and understand behavioral control, and to gain insights into the neurophysiological aspects of psychiatric disorders. By computationally studying how emotion relates to information processing and reinforcement we hope to extend the analogy between RL and behavior.

To study the low-level regulatory influence of emotion on information processing, we use a computational RL model. Besides RL, our approach is based upon the following hypotheses. **1.**) The *Simulation Hypothesis*, which assumes

that thinking is internal simulation of behavior using the same sensory-motor systems as those used for overt behavior [Hesslow, 2002] **2.**) *interactivism*, stating that thinking emerges from continuous interaction with the environment [Bickhard, 2001].

These hypotheses have several important characteristics in common [Broekens, 2005b], amongst which the following are particularly important for this paper:

a.) These hypotheses are primarily about neuronal systems, but do allow connectionist but non-neuronal modeling, the basis of our model.

b.) Emotion plays a role in information processing.

c.) These hypotheses closely relate to Damasio's [1994] concept of thinking as an "as-if body loop", involving simulated actions that are evaluated by their *somatic markers*, emotional impact estimators. Four systems are critically involved: the body; the somato-sensory cortex (SSC), the emotional marker system that receives information from the body; the sensory and association cortexes (SC/AC); and the ventromedial prefrontal cortex (VM-PFC), the system that stores relations between factual representations stored in the SC/AC and somatic markers stored in the SSC. Interaction with the environment enables the VM-PFC to learn these links. Two important processing mechanisms are the "body-loop" and the "as-if body loop". When facts about a situation are recognized, the SC/AC activate the VM-SSC, and links between the situational facts and emotional outcomes are activated. In the "body-loop", the VM-SSC activates the body, and the SSC that stores somatic-markers is organized according to the body. This loop thus involves the emotional evaluation of action. In the "as-if body loop", the VM-PFC signals the SSC to reorganized itself directly without signaling the body. This loop thus involves the emotional evaluation of simulated action. The "as if" loop produces imagined future factual-emotional states, and the somatic marker part of such states is the state's predicted accumulative emotional outcome (reward/punishment). This marker signal is used to bias decision-making [Damasio, 1994]. Even though we do not model the body of the agent, we use the somatic marker concept to understand the relation between reinforcement learning (RL), emotion and decision-making.

In this paper we first introduce our computational approach without emotional feedback. Next, we introduce our concept of emotion and pleasure in more detail, and we explain how pleasure is used to control the amount of anticipatory simulation of the agent. Finally, we discuss our results, related work and give directions for future research.

2 Computational Approach

Our experiments are performed in a gridworld, a two-dimensional grid with positively and negatively reinforced locations, in our case, lava (negative reinforcement of –1), roadblocks (–0.5), food (+1.0) and empty cells (Figure 1). The agent can move everywhere, but is discouraged to walk on the lava (by a negative reinforcement). The agent's perceptual field has either a chessboard, 8 neighbor (Figure 1b), or a cityblock, 4 neighbor metric (Figure 1a, c). In, e.g., Figure 1c, the agent would perceive "eleee" representing the (l)ava left of the agent and the (e)empty cells above, right, beneath, and below the agent.

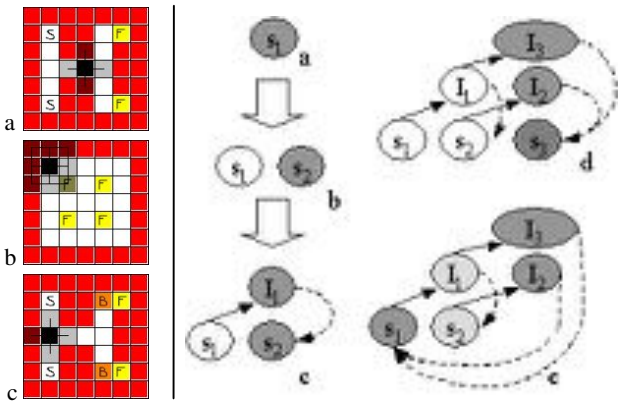


Figure 1 (left) and 2 (right). Fig. 1: three different experimental settings: agent (black), lava (dark gray, red), possible food (*F*), roadblock (*B*), possible start location (*S*). Tasks from left to right: *find food*, *forage*, *invest*. Fig. 2: examples of the agent's memory.

2.1 Hierarchical-State Reinforcement Learning

We first explain the basic model without emotional feedback. The agent's memory structure is modelled by a directed graph. The memory is adapted while the agent interacts with its environment (online learning) in the following way. The agent selects an action, $a \in A$, from its set of potential actions $A=\{u, d, l, r\}$, executes the action in the grid-world and perceives the result of that action, p. This is combined into a *situation*, $s=<a, p>$, that is stored in the agent's memory according to a basic rule: *if a situation s occurs, the agent creates a node in the graph if and only if there does not exist a node for s*. For example in Figure 1c, if the agent has moved down, "d", and perceives "eleee". In an initially empty model a node is created to represent the situation $s_1=<d,eleee>$ (Figure 2a), because the graph does not yet contains this node. Now the agent moves again, resulting in a new situation, e.g., $s_2=<d,elele>$, resulting in a new node that represents s_2 (Figure 2b). To model that s_2 follows s_1 (or s_1 predicts s_2), the previous situation, s_1, is now connected to the current situation, s_2, by creating a new node, an *interactron,* between s_1 and s_2 with edges as shown in Figure 2c. This process continues, never violating the basic rule. Also, the process is recursively applied to active interactrons. Active in this case means that an interactron corresponds to the history of observed situations, e.g., node I_1 in Figure 2c. If situation s_2 is followed by s_3, the resulting memory structure is shown in Figure 2d, with active nodes s_3, I_2 and I_3. If, on the other hand s_2 is followed by s_1, the resulting structure is shown in Figure 2e, with active nodes s_1, I_2 and I_3.

If at a later time the sequence of situations s_1s_2 is again observed then, according to the rule, I_1 is not created again. Instead, a counter υ, the *usage* of interactron I_1, that is initially zero is increased by one. This υ can be used to calculate the probability $P(s_2 \mid s_1)$ using the following more generic formula:

$$P(x \mid y) = \upsilon_x \Big/ \sum_{i=1}^{|X_y|} \upsilon_{x_i}$$

,where y is an active interactron or situation, $x \in X_y = \{x_1,\ldots,x_n\}$ the set of predicted situations by y (represented by their corresponding interactrons, e.g., I_1 representing the prediction of s_2). This formula is true, for I_1 is conditionally active upon s_1, and υ is only increased if an interactron is active and multiple sequences other than s_1s_2, e.g., s_1s_3, s_1s_4 etc., have their own interactron attached to s_1 with its own υ increased if and only if the corresponding sequence is observed. Furthermore, we define a threshold, θ, representing the minimal "survival probability" for an interactron. If $P(x \mid y) < \theta$, the corresponding interactron is forgotten and removed from the memory, including its dependencies. This corresponds to Bickhards [2000] notion of interaction (de)stability based on consistent confirmation of predicted interactions, see also [Broekens and DeGroot, 2004].

The memory maintains a distributed, hierarchical prediction of the next situation. Every active interactron predicts potential next situations, k of these interactrons can be active, and the 1st till k-th interactron predict potential next situations with a history of length 1 till k respectively (e.g., I_3 is a $k=2$ interactron with history s_1s_2). Learning in the context of this memory can be seen as the online learning of 1…k-th order Markov Decision Processes in parallel.

In addition to a predictive probability, every interactron has a reinforcement value, called a *marker*, μ, with $\mu = \lambda + \nu$, where λ is the interactron's *direct reinforcement* value and ν is a back-propagated *indirect reinforcement* value. Thus, the value of an interactron is a function of it's own reward and the rewards of those situations it predicts. More specific, first, all k active interactrons are reinforced by a signal from the environment, r^t, at time t. For every such interactron y, λ_y is adapted according to the formula:

$$\lambda^{t+1}{}_y = \lambda^t{}_y + (r^t - \lambda^t{}_y) \times \rho$$

, where ρ is the agent's learning rate. Second, for every interactron y, ν_y, is calculated as follows:

$$\nu^{t+1}{}_y = \sum_{i=1}^{|X_y|} \mu^t(x_i \mid y) \times P(x_i \mid y)$$

, where $\mu^t(x_i \mid y)$ is defined as the marker of interactron x_i, with x_i predicted by y. This indirect part of an interactron's (say y) value is thus the weighted average of the markers belonging to the interactrons X_y that represent the situations that y predicts, where weighted is according to the probability distribution $P(x_i \mid y)$ over all i.

Action-selection is based on the parallel inhibition and exhibition of actions in the set of actions, A. The inhibition/exhibition originates from the k active interactrons and is calculated using the formula:

$$l^t(a_h) = \sum_{i=1}^{k} \sum_{j=1}^{|X_{y_i}|}{}^{*} \mu^t(x^i{}_j \mid y_i) \times P(x^i{}_j \mid y_i)$$

, where $l^t(a_h)$ is defined as the level of activation of an action $a_h \in A$ at time t, y_i an active interactron, and x^i_j predicts action a_h. This last clause is needed, for the memory stores action-perception pairs and any of these pairs that are predicted by any of the k active interactrons should inhibit (negative marker) or exhibit (positive marker) the corresponding action, but not other actions. Additionally, of all good actions (any $l^t(a_h) > 0$) the best action a_h, i.e., $l^t(a_h) = max(l(a_1),\ldots,l(a_{|A|}))$, is always selected. If there are only bad actions (all $l^t(a_h) < 0$) a stochastic selection is made based on $(l(a_1),\ldots,l(a_{|A|})$; the action with the highest activation therefore has the highest chance of being chosen resulting in a probabilistic Winner-Take-All action-selection.

The process described in this section is our agent's "body loop". Next, we describe our agent's "as-if" loop, its simulation mechanism. For a discussion on the relation between Damasio's somatic marker hypothesis and our computational model, see [Broekens, 2005b].

2.2 Internal Simulation and Action-Selection Bias

To study anticipatory simulation we add the following capability to our model: after every real interaction with the environment, the model simulates one time-step ahead. Instead of selecting an action based on past interactions the following process is executed:

1.) *Interaction-selection*: at time t select a subset of to-be-simulated interactions from the set of interactions predicted by all k active interactrons.

2.) *Simulate*: send the subset of selected interactions to the model as if they were real interactions. The memory advances to time $t+1$.

3.) *Reset-state*: to be able to select an appropriate action, reset the memory's state (the active interactrons) to the previous timestep, i.e., time t.

4.) *Action-selection*: select the next action using the standard mechanism described above. Thus, the propagated markers of the simulated predicted interactions directly bias action-selection. Our anticipation mechanism is best understood as *state anticipation* [Butz *et al*, 2003].

5.) *Reset-markers*: reset μ, λ and ν of the interactions that were changed at step 2 (simulation) to the values of μ, λ and ν of these interactions before step 2.

Step 1 selects predicted interactions to be simulated, and is a critical component in our simulation mechanisms since it defines the amount of internally simulated information. In a previous experiment [Broekens, 2005] we used four static selection criteria (also referred to as *simulation strategies*).

a.) No simulation (NON). The actions are selected as described in the previous section and the 5-step simulation procedure is not executed. **b.**) Simulation of the predicted best interaction (BEST). The winning interaction of the WTA selection resulting from step 1 is sent to the model for simulation (step 2). Any real interaction is accompanied by a reinforcement signal. As this is a simulation we lack such a signal. Instead, this signal is simulated using the μ of the winning interaction as reinforcement. We simulate the predicted interaction and its associated value. **c.**) A selection of the predicted 50% best interactions, i.e., a more balanced selection, (BEST50). Again we simulate the reinforcement signal using the μ's of the simulated interactions. **d.)** All of the predicted interactions (ALL).

In essence, NON, BEST, BEST50 and ALL simulate different values for the *selection threshold* of the WTA interaction selection ranging from infinite (NON) to high (BEST) to medium (BEST50) to low (ALL). This threshold filters the set of predicted interactions used to simulate. The final result of simulation is a bias to the predicted rewards of the set of next possible interactions, with action-selection based on these biased rewards (Figure 3). This means that our model of internal simulation influences action-selection in a way that is compatible with the somatic marker hypothesis [Damasio, 1994] and the simulation hypothesis [Hesslow, 2002]. For more on the compatibility between our model and the simulation hypothesis see [Broekens, 2005].

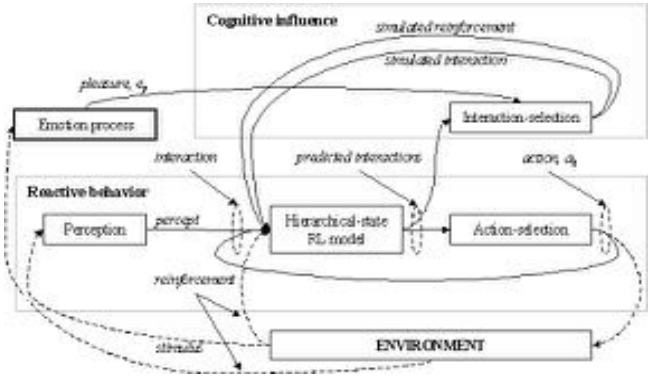


Figure 3. Architecture of the different components in our model.

2.3 Differences in Performance of Static Simulation Strategies Motivate the Feedback Control

In a previous study [Broekens, 2005] we showed that simulation in general, and simulation of all possible next interactions (ALL) in particular, has a clear adaptive advantage. The agent learns the tasks quicker and converges better to the solution. The agent had to learn three tasks (Figure 1):

1.) Continuously try to *find* a randomly changing food location, thereby learning the optimal route to both possible food locations in the gridworld maze (Figure 1a).

2.) Learn to *forage* (Figure 1b). Now, the agent is initially placed in the environment, after which it should explore and find food. Again, food locations are randomly selected.

3.) The same as the first, but the agent additionally had to learn to accept an initial negative reinforcement (roadblock in Figure 1c) in order to get to a larger positive one (food in Figure 1c). With this task we wanted to test how the different simulation strategies handle *investment*, which is a relevant problem for natural adaptive agents [Doya, 2002].

Intuitively it is not really a surprise that ALL "wins", as it is the heuristic using the most information. However, for some experimental settings BEST or BEST50 do result in a better performance (i.e., a smaller amount of simulation results in a better performance). This suggested a relation between the parameters of the experimental setting, and the effect of the amount of simulation used by the agent.

Analysis of this relation revealed that the *goal orientedness* of the task and the *complexity of the task* influence this performance. When the agent is solving a goal oriented task (*find food*, *invest*), it benefits from a narrow (i.e., BEST) simulation strategy with a high learning rate, while in an uncertain or more exploratory task (*forage*) it benefits from a broad (i.e., BEST50 or ALL) simulation strategy.

Simple goal-oriented tasks are solved by quickly propagating the delayed reward to the beginning, specifically if there is "just one hill to climb". Local solutions converge to a global solution. The faster the convergence the quicker the global solution is found, as reflected by previous results.

If a task is complex, the agent benefits from broader simulation, for this allows it to mentally explore multiple options and make a more balanced choice. This relates to the exploration-exploitation problem [cf. Doya, 2002]. Essentially our agent has to vary its *simulation strategy* (instead of its action selection) between mental exploitation and mental exploration.

These findings suggested that it is beneficial to the agent to dynamically adapt simulation to accommodate the task. Additionally, we hypothesized that dynamic adaptation of simulation could outperform any of the four static strategies tested, for dynamic adaptation could be beneficial to the agent at *different stages of learning a task*. The main hypothesis addressed in this paper is: *if action-selection-bias is induced by an amount of simulated anticipatory behavior, and if this amount is dynamically controlled by pleasure feedback, then this provides additional survival value to an agent, compared to a static amount of simulation*. Our approach is compatible with Cañamero's [2000] view on why and how emotion systems should be designed.

3 Emotion as Pleasure and Arousal Factors That Control Information Processing

Before describing how we add emotional feedback to the simulation mechanism, we present some rationale for our concept of emotion. Emotion influences thinking. This influence is found at low and high levels of information processing and is both positive as well as negative. For example, at the neurological level malfunction of certain brain areas not only destroys or diminishes the capacity to have (or express) certain emotions but also has the same effect on the capacity to make sound decisions [Damasio, 1994] and on the capacity to learn new behavior [Berridge, 2003], which indicates that these areas are linked to emotions as well as "classical" cognitive and instrumental learning phenomena. At the cognitive psychological level a person's beliefs about something are updated according to the emotion. The current emotion is used as information about the perceived object [Clore and Gasper, 2000; Forgas, 2000], and emotion is used to make the belief resistant to change [Frijda and Mesquita, 2000]. Emotions are "at the heart of what beliefs are about" [Frijda *et al.*, 2000]. For example, your belief about roller coasters tells you something about the emotion attached to your cumulative experiences with roller coasters.

More specifically, emotion is related to the regulation of adaptive behavior and to information processing. Emotions can be defined as states elicited by rewards and punishments [Rolls, 2000]. Behavioral evidence suggests that the ability

to have sensations of pleasure and pain is highly connected to basic mechanisms of learning and decision-making [Berridge, 1998; Cohen and Blum, 2002]. Behavioral neuroscience teaches us that positive emotions reinforce behavior while negative emotions extinct behavior, so at this lower level one type of regulation of behavior has already been established—i.e., approach versus avoidance. The emotion resulting from an unconditioned natural stimulus is associated with the conditioned stimulus or with a specific action. In the future, upon presentation of the conditioned stimulus to the animal, this association results either in more actively *choosing the action* that leads to the unconditioned stimulus (rats' lever pressing behavior) or in *behavior that is associated* with the unconditioned stimulus (Pavlov's dog producing saliva). At this lower level, emotion has a direct—mostly associative—effect (but also other effects are reported [Dayan and Balleine, 2002]).

At the higher level of cognitive psychology, evidence suggests that the processes involved in emotion are crucial for both evaluating the world around us at different levels of abstraction [Scherer, 2001] as well as actually taking action [Frijda, 2000]. Emotion also plays a role in the regulation of cognitive processes. Scherer [2001] argues that emotions are related to the continuous checking of the environment for important stimuli. More resources are allocated to further evaluate the implications of an event, only if the stimulus appears important. This suggests that certain emotions are related to regulation of the *amount of information processing*. This finding provides an important clue to our approach of adding emotional control to the amount of simulation used by the agent. Furthermore, in the work of Forgas [2000] the relation between emotion and information processing strategy is explicit: depending on the strategy used, the influence of mood on thinking changes.

Although many different emotions (and emotion theories) exist, and emotion consists of many different components—e.g., facial expression, a tendency to act, subjective evaluation of the situation—, the *core-affect* theory of emotion states that emotion (mood) consists of two fundamental factors, *pleasure* and *arousal* [Russell, 2003]. Pleasure relates to emotional valence, while arousal relates to action-readiness, or activity, of the organism. Many different situations can be emotionally described using these two factors, for example, winning the lottery (a high arousal high pleasure emotion), or losing a friend (a low arousal and low pleasure emotion). Although Mehrabian [1996] argues for dominance as a third factor, he agrees with, and shows considerable evidence for, the pleasure and arousal factors.

Certain cognitive appraisal theories argue that pleasure and arousal can be produced by very simple stimulus checking functions. This suggests that low-level mechanisms like intrinsic pleasantness checks and suddenness checks are involved [Scherer, 2000].

The suggestion that pleasure and arousal factors are fundamental to emotion, that these factors can be produced by simple mechanisms and that these factors can influence further information processing inspired us to look at how these two factors could result from low-level features of the agent's memory structure and its performance, and subsequently how these factors could then influence information processing in a way that is compatible with cognitive appraisal theory. In this paper we focus on the pleasure factor.

3.1 Pleasure As a Measure for Relative Task-Performance

According to cognitive appraisal theory positive emotions are related to top-down goal oriented processing while negative emotions are related to bottom-up stimulus oriented processing [Fiedler and Bless, 2000]. Furthermore, emotion is often seen as an indication of the current performance of the agent [Clore and Gasper, 2000]. To capture these findings we measure pleasure in the following way:

$$e_p = (\bar{r}_{star} - (\bar{r}_{ltar} - f\sigma_{ltar}))/2f\sigma_{ltar}$$

The current pleasure, e_p, of the agent is the short-term running average over the reinforcement signal, r, with a window size of *star* steps, normalized around the agent's long-term running average over the same reinforcement signal with a window size of *ltar* steps. This value is normalized using f times the standard deviation of the long-term distribution of reinforcement signals σ_{ltar}. So, e_p is a continuous measure for how well the agent is currently performing on a task, relative to what it is used to, according to the recent past. A large f results in smaller fluctuations around 0.5, while a small f results in larger fluctuations around 0.5. Also, e_p is clipped between 0 and 1. Information processing can be influenced by e_p in the following way (Figure 3). When $e_p=1$, interaction-selection (Step 1) selects only the best interactions for simulation, i.e. a high selection threshold. When $e_p=0$ it selects all interactions, i.e., a low selection threshold. The agent thus varies between BEST and ALL depending on its pleasure. It can be argued that our use of pleasure relates more to mood than to emotion, due to its timescale. Moods typically occur at longer timescales, while emotions are short complex reactions to events. Pleasure in our case is measured over multiple interactions and does not react to one interaction in particular. Even if e_p is interpreted as the agent's mood, the modeled effects of positive versus negative emotion is consistent with the previously mentioned ideas about top-down versus bottom-up processing related to respectively positive and negative emotions as well as to the concept of emotion influencing the amount of processing needed. If the agent goes well, little processing (focussed attention) is needed, if it goes bad more processing (broad attention) is needed.

4 Experimental Setup

To test our hypothesis we created a combined task in which simple and complex elements are present as well as goal oriented and exploratory behavior is needed. The first half consists of the *find food* task (Figure 1a), and the second half consists of the *invest* task (Figure 1c). The agent is unaware of this change; it is abruptly replaced in a slightly different environment and has to learn about this change by interacting with the environment. The hypothesized effect is that the agent dynamically adapts the amount of simulation

according to the change in complexity and goal orientedness. We predicted the following changes to simulation during the task: BEST→ALL→BEST. BEST performs best on the goal-oriented *find food* task. The change to the *invest* task induces a pleasure decrease, resulting in simulation close to ALL: mentally explore the new task. During learning of the *invest* task, simulation should return to one that is close to BEST because the agent's pleasure increases, resulting in goal oriented behavior of the agent.

f:	*1*		*1.5*		*2*	
star:	*50*	*100*	*50*	*100*	*50*	*100*
ltar:	200	400	200	400	200	400
	250	500	250	500	250	500
	375	750	375	750	375	750
	500	1000	500	1000	500	1000
	750	1500	750	1500	750	1500

Table 1: *ltar*, *star*, and *f* configurations used in the experiment.

One experimental setting is a combination of *f*, *star*, *ltar*, θ and ρ. These parameters are varied as follows: the forgetting rate θ*=(0, 0.01, 0.03, 0.05)*, learning rate ρ*=(1, 0.8)* and *ltar*, *star* and *f* according to Table 1. For every experimental setting the agent had 255 *trials* (defined as one *run*) to get to the food. It had to learn the task within these 255 trials, which showed to be enough to conclude convergence.

For every experimental setting, we recorded the agent's total number of actions needed to complete a run (i.e. 255 trials), and averaged over 15 runs. This resulted in averages for *5×6=30* (*f*, *star*, *ltar*) configurations per (θ, ρ) configuration. The goal of these experiments is not to find out what the exact parameters are to get the best dynamic result, but to investigate the potential benefit of pleasure controlling simulation effort in general. We assume that there should be an *overall benefit* to emotional feedback. Therefore, averaging again aggregates these 30 averages. The result is one value per (θ, ρ) depicted by the red (gray) lines in Figure 4. Red lines should be interpreted as the average performance of an agent that uses emotional feedback to dynamically control the amount of simulation (DYN). Performance is in terms of the total *number of interactions* needed to complete a run (Figure 4a and c), and *mental effort* in terms of the total number of simulated interactions needed to complete a run (Figure 4b and d). Black lines show the corresponding performance of the static strategies (NON, BEST, BEST50, ALL) averaged over 30 runs per (θ, ρ) configuration.

5 Results

The performance of our dynamically adapting agent is comparable to (Figure 4a and c), and in several special cases even better than (Figure 4e, result of one setting averaged over 30 runs instead of 15), the performance of our static agents. If this effect is put in light of total simulation (mental) effort, it is even more dramatic. DYN uses about 33% of the mental effort needed for ALL and about 70% of the effort needed for BEST50 but performs comparably. The predicted effect of the pleasure feedback is confirmed. Figure 5 depicts a typical pleasure flow (15 run e_p average) of an agent that uses DYN. Just after the task switch (at trial 128) a steep decrease of pleasure is observed, this results in more simulated interactions, i.e., broader attention. While exploring, the agent improves at the *invest* task, and pleasure gradually increases, resulting in goal-directed simulation.

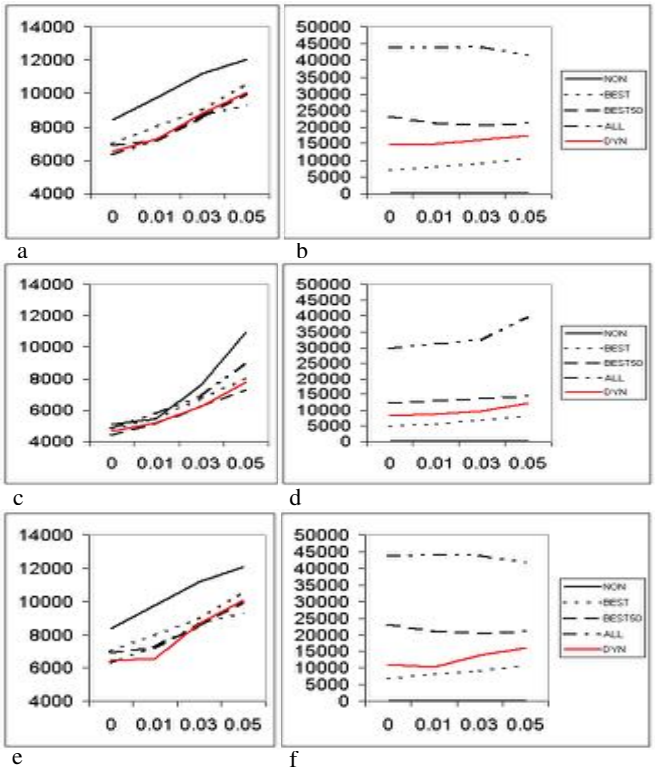


Figure 4. Figure 4a and b ρ=0.8, 4c and 4d ρ*=1.* Figure 4e,f, DYN (*star=100, ltar=1500, f=1*) performing better (one-tailed *t*-test, *n*=30, α=0.05) than static strategies with ρ*=0.8 and* θ*=0.01.*

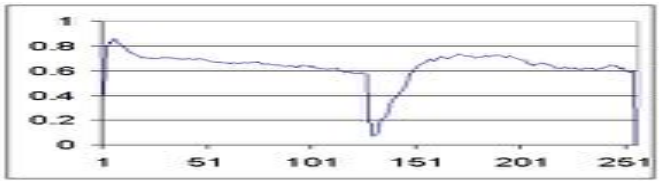


Figure 5. Pleasure flow during one run, averaged over 15 runs.

6 Discussion and Conclusions

Under the assumption that total simulation effort positively correlates with total energy consumption of the agent, decrease of mental effort reduces the energy need for information processing, thereby saving energy for occupancies other than foraging. If dynamic adaptation reduces mental effort and if this is an hereditary feature, it becomes evolutionary advantageous. This suggests that dynamic adaptation of the amount of simulation has a strong evolutionary drive.

Our results show that the relation between (1) positive emotions and top-down goal oriented thinking, and (2) negative emotions and bottom-up stimulus driven thinking could result from the feedback of a simple measurement of

the performance of the agent to the selection threshold of the simulation mechanism. These results show one possible relation between emotion and the simulation hypothesis, as well as provide experimental evidence for the fact that even simple emotional integration processes can be used to adapt cognitive processes.

6.1 Related Work

Our work is highly related to Gandanho's [2003] work on the "Alec" architecture. However, in their RL based adaptive system, stochastic action-selection is biased by a fixed value produced by a rule-based cognitive system. In our system this value is dependent on the predicted states and the cognitive process is not separated from the adaptive system. We chose not to separate the cognitive system from the reactive system, as this is important for the evolutionary continuity between simulating and non-simulating agents [Broekens, 2005; Cruse, 2002; Hesslow, 2002].

The "Salt" model by Botelho and Coelho [1998] relates to ours in the sense that the agent's effort to search for a solution in its memory depends on, among other parameters, the agent's mood valence. Our approach differs in that we focus on simulation of behavior (not specifically targeted towards search), we use a dynamic influence to link emotion to the cognitive system (not a rule-based system), and we specifically define how our agent's mood is produced.

Our work relates to emotion and motivation based control/action-selection, in that it explicitly defines a role for emotion in biasing behavior-selection [Avila-Garcia and Canamero, 2004; Canamero, 1997; Velasquez, 1998]. The main difference is that in these studies emotion directly influences action-selection (or motivation(al states)), while we have studied the indirect effect of emotion as a metalearning parameter affecting information processing that on its turn influences action-selection (cf. Gandanho [2003]).

Up until now our agent is unable to learn the representation of a goal (what is a goal) and thus is unable to consider different goals in its final action selection. We learn from behavioral neuroscience that rats adapt learned behavior contingent on their drives (i.e., lever-pressing when hungry versus button-pushing when thirsty) [Dayan and Balleine, 2002]. They argue that the rat's motivation acts as a gate between the learned predictive state and the incentive value associated with it. Such a mechanism can be implemented using a Markov Decision Process [Smith *et al,* 2003]. They model a conditioning task whereby the learned reward is multiplied by an artificially varied "gating factor", i.e., a simulated dopamine signal that is necessary for the agent to see the consequences of its actions.

However, implementations such as [Smith *et al*, 2003] are still limited since many animals develop multiple complex goals, suggesting that they can learn to use many representations as gating factor for the predicted reinforcement signal in a certain situation. In this case, a learned goal can influence behavior without the behavior being directly associated with a positive or negative reinforcement signal. Learned goals could even become reinforcers by themselves. This approach relates to one proposed by Singh *et al.* [2004], where multiple different reinforcement techniques are used to learn hierarchical collections of skills that function as intrinsically motivating actions for the agent. Further, it relates to work by Gadanho [2003], where multiple goals—related to homeostatic variables—determine the reinforcement for the adaptive system, and to work on emotion learning by, for example, Bothelo and Coelho.

6.2 Future work

We have investigated one way in which pleasure can influences information processing. Combining *arousal and pleasure* as feedback to control simulation might give additional insights into the relation between these two factors, as well as introduce a second learning metaparameter.

To measure arousal, the agent could compare to what extend the predicted environment equals the actual environment. This measurement is called the stimulus predictability check [Scherer, 2000]. We can implement this in our model by comparing the probabilities of next interactions with the actually occurring interactions.

Another way to measure arousal is the stimulus familiarity check [Scherer, 2000]. This check measures how much of the environment is actually known. In our model we can count the number of active interactions in the state hierarchy (high number = familiar, low number = unfamiliar).

These two arousal measurements can be integrated into one signal, say e_a that, e.g., influences the absolute amount of effort put in simulation (information processing). A high e_a results in a large amount of effort put into simulation, while a low e_a results in a low amount of effort. The e_a factor combined with e_p results in a distribution of maximum available simulation steps over the potential next interactions. Along these lines, we plan to adapt our model so that it is able to simulate multiple steps ahead depending on a cut-off depth based on the total amount of effort available for that specific branch. This approach is highly similar to planning and algorithms for depth-first, breadth-first and iterative deepening search. We hope that techniques following from our research are generic in terms of their ability to modify solution-search behavior in these kinds of algorithms.

A different way to influence simulation is by letting e_a control the amount of randomness in the interaction selection process. This is analogous to the role of noradrenaline as proposed by Doya [2002].

6.3 Conclusions

Experimental results show that if pleasure is used to dynamically adapt the amount of simulation, this results in a learning performance that, at least, equals static simulation strategies. Importantly, our results show a major decrease of mental effort required for this performance. This observation is relevant to the understanding of the evolutionary plausibility of the simulation hypothesis, as increased adaptation at lower cost is an evolutionary advantageous feature. In addition, our results provide clues of a relation between the simulation hypothesis and emotion theory.

Acknowledgements

We thank Jeroen Eggermont for useful discussions, and the MNAS reviewers and Walter Kosters for useful suggestions.

References

[Avila-Garcia and Cañamero, 2004]. O. Avila-Garcia and L. Cañamero. Using hormonal feedback to modulate action selection in a competitive scenario. In: *From Animals to Animats 8: Proc. 8th Intl. Conf. on Simulation of Adaptive Behavior.* MIT Press, Cambridge, Massachusetts.

[Berridge, 2003] K. C. Berridge. Pleasures of the brain. *Brain and Cognition 52*.

[Botelho and Coehlo, 1998]. L. M. Botelho and H. Coelho. Information processing, motivation and decision making. In: *Proc. 4th International Workshop on Artificial Intelligence in Economics and Management*.

[Butz *et al*., 2003] M. V. Butz, O. Sigaud and P. Gerard. Internal models and anticipations in adaptive learning systems. In: *Anticipatory Behavior in Adaptive Learning Systems.* Springer (LNAI 2684).

[Bickhard, 2000] M. H. Bickhard. Motivation and emotion: an interactive process model. In: *The Caldron of Consciousness*. John Benjamins, New York.

[Broekens and DeGroot, 2004] J. Broekens and D. DeGroot. Emergent Representations and Reasoning in Adaptive Agents. In: *Proc. ICMLA'04*. IEEE.

[Broekens, 2005] J. Broekens. Internal simulation of behavior has an adaptive advantage. In: *Proc. CogSci'05*. (in press).

[Broekens, 2005b] J. Broekens. Computational Investigations of the Regulative Role of Pleasure in Adaptive Behavior. TR 2005-06, LIACS, Leiden University. http://www.liacs.nl/~broekens/BroekensTR2005-06.pdf.

[Cañamero, 2000]. D. Cañamero. Designing emotions for activity selection. *Dept. of Computer Science Technical Report DAIMI PB 545*. University of Aarhus, Denmark.

[Clore and Gasper, 2000] G. L. Clore and K. Gasper. Feeling is believing: some affective influences on belief. In: *Emotions and Beliefs,* Cambridge Univ. Press, Cambridge, UK.

[Cohen and Blum, 2002] Jonathan D. Cohen and Kenneth I. Blum. Reward and decision. *Neuron 36*.

[Cruse, 2002] H. Cruse. The evolution of cognition: a hypothesis. *Cognitive Science 27.*

[Damasio, 1994] A. R. Damasio. *Descartes' error: Emotion, reason, and the human brain.* G.P. Putnam, New York.

[Dayan and Balleine, 2002] P. Dayan and B. W. Balleine. Reward, motivation, and reinforcement learning. *Neuron 36*.

[Davidson, 2000] R. J. Davidson. Cognitive neuroscience needs affective neuroscience (and Vice Versa). *Brain and Cognition 42*.

[Doya, 2000] K. Doya. Metalearning, neuromodulation, and emotion. In: *Affective Minds*. Elsevier Science B.V.

[Doya, 2002] K. Doya. Metalearning and neuromodulation. *Neural Networks 15.*

[Fiedler and Bless, 2000] K. Fiedler and H. Bless. The formation of beliefs at the interface of affective and cognitive processes. In: *Emotions and Beliefs.* Cambridge Univ. Press, Cambridge, UK.

[Forgas, 2000] J. P. Forgas. Feeling is believing? The role of processing strategies in mediating affective influences in beliefs. In: *Emotions and Beliefs.* Cambridge University Press, Cambridge, UK.

[Frijda and Mesquita, 2000] N. H. Frijda and B. Mesquita. Beliefs through emotions. In: *Emotions and Beliefs.* Cambridge Univ. Press, Cambridge, UK.

[Frijda *et al*., 2000] N. H. Frijda, A. S. R. Manstead and S. Bem. The influence of emotions on beliefs. In: *Emotions and Beliefs.* Cambridge Univ. Press, Cambridge, UK.

[Gadanho, 2003] S. C. Gadanho. Learning behavior-selection by emotions and cognition in a multi-goal robot task. *Journal of Machine Learning Research 4*.

[Griffith, 1999] P. E. Griffith. Modularity & the psychoevolutionary theory of emotion. *Mind and Cognition: An Anthology*

[Hesslow, 2002] G. Hesslow. Conscious thought as simulation of behaviour and perception. *TICS 6*.

[Mehrabian, 1996] A. Mehrabian. Framework for a comprehensive description and measurement of emotional states. *Gen., Soc. and General Psych. Monographs 121*.

[Montague *et al.*, 2004] P. R. Montague, S. E Hyman and J. D. Cohen. Computational roles for dopamine in behavioural control. *Nature 431*.

[Rolls, 2000] E. T. Rolls. Precis of The brain and emotion. *Behavioral and Brain Sciences 23*.

[Russell, 2003] J. A. Russell. Core affect and the psychological construction of emotion. *Psychological Rev. 110.*

[Scherer, 2001] K. R. Scherer. Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, Methods, Research*. Oxford Univ. Press, New York.

[Smit *et al.*, 2003] A. Smith, S. Becker and S. Kapur. From dopamine to psychosis: a computational approach. In *Proc. KES'03*. Springer (LNAI 2773).

[Singh *et al.*, 2004] S. Singh, A. G. Barto and N. Chentanez. Intrinsically motivated reinforcement learning. *Proc. NIPS'04*. MIT Press, Cambridge, Massachusetts.

[Sutton and Barto, 1996] R. S. Sutton and A. G. Barto. *Reinforcement Learning: an introduction*. MIT Press, Cambridge, Massachusetts.

[Velasquez, 1998]. J. D. Velasquez. A computational framework for emotion-based control. In: *SAB'98 Workshop on Grounding Emotions in Adaptive Systems*.

ROUTINE ACTION: COMBINING FAMILIARITY AND GOAL ORIENTEDNESS

Nicolas Ruh, Richard P. Cooper and Denis Mareschal
School of Psychology, Birkbeck, University of London
Malet Street, London, WC1E 7HX, UK.
n.ruh@psychology.bbk.ac.uk, r.cooper@bbk.ac.uk, d.mareschal@bbk.ac.uk

Abstract

Two current approaches to modelling naturalistic sequential routine action selection differ along two dimensions: (a) the number of systems required and (b) the nature of the underlying task representation. We present findings from a study that supports a combination of the two computational accounts, namely a familiarity-dependent basic system, interfaced with a higher-level supervisory system to bias it at crucial points in a sequence. In order to elaborate this position, we explore a connectionist reinforcement model of routine action that (a) learns goal-directed action sequences through a combination of exploration and exploitation, and (b) offers the prospect of being interfaced with a supervisory system.

1 Introduction

The sequencing of actions in routine tasks, such as dressing or preparing tea or coffee, is prone to slips and lapses. In a series of diary studies, Reason [1984] found that such slips and lapses were particularly frequent when the actor was fatigued or distracted. The standard account of this finding is offered by the dual systems view of action control proposed by Norman & Shallice [1986]. According to this view, action is controlled by the Contention Scheduling system (CS), which consists of schemas that compete for selection. This system is prone to error and its functioning can be "captured" by bottom-up input, but equally it may be biased in a top-down fashion by a second system, the Supervisory Attentional System (SAS) that acts to prevent errors (by exciting appropriate schemas and inhibiting inappropriate ones) and to guide deliberate behaviour (by selectively exciting appropriate schemas in sequence). Within the dual systems view, routine sequential behaviour can be controlled by CS in the absence of input from the SAS, provided the task and situational context are fully routine, but slips and lapses may arise if the task and context are not fully routine, and the SAS is not engaged.

This view is exemplified in a computational model of CS that employs hierarchically organized Interactive Activation Networks (IAN) in which symbolically represented schemas compete for selection. The model can account for slips and lapses of routine action, as well as more flagrant errors of action shown by patients with frontal brain damage [Cooper & Shallice, 2000].

In contrast, Botvinick & Plaut [2004] claim to capture both the normal and the neurological data with a single embedded Simple Recurrent Network (SRN). The SRN learns to reproduce a corpus of well-ordered behaviour, but when noise of varying levels is added to units in the hidden layer, the SRN reproduces errors that range from minor slips and lapses to more serious disorganisation of action. An important feature of the SRN account is that schemas are not explicitly represented. Rather, they emerge as continuous attractors within the state space of the SRN. These task representations "overlap structurally, sharing graded, multidimensional similarity relations" [Botvinick & Plaut, 2002, p. 299]. This, they argue, overcomes a problem of more traditional approaches that require one discrete representation/schema for each version of the task.

The SRN model has some further appealing features. Most significantly, it addresses the issue of learning – something that is not addressed by the IAN model. However, the supervised learning regime employed by Botvinick & Plaut [2004] is implausible because it employs an explicit error signal that is rarely available. It is furthermore heavily dependent on the exact composition of the training set. If that set is not balanced in a specific way, the SRN will not be able to reproduce all sequences in the training set [see Ruh *et al.*, to appear, for further details].

A second criticism of the SRN model concerns its treatment of goals. Botvinick & Plaut [2004] employ instruction units (such as *make-coffee* and *make-tea*) to coerce their model into the production of different learnt action sequences, but they deny that these units encode or represent goals. In our view these units do represent goals, but the representation is too impoverished for the control of sequential action in anything but routine situations. They cannot, for example, aid in detection and recovery from errors.

Impoverished goal representations would not necessarily be a problem for the SRN model if it was conceived of as working in conjunction with a supervisory system to control behaviour in non-routine situations (although if that were the case, interfacing the SRN model in its present form with the SAS would present difficulties). Botvinick & Plaut [2004] appear, however, to also deny this.

The Botvinick & Plaut model therefore differs from the Cooper & Shallice model in two key respects: its representation of schemas and its appeal to a single system for the control of action. It is important to note that these differences are independent. In the next section we briefly present our findings in a recent study that investigates sequential routine behaviour of normal subjects. These findings support the dual-systems view, but suggest as well that the basic level system is gradually adjusted to familiarity. This feature is naturally captured within an SRN model. In the final section we therefore present a tentative exploration of a connectionist reinforcement model that is aimed at combining the respective strengths of the two existing computational accounts and that offers the prospect of accommodating our data.

2 New data

Our findings in a recent sequential routine action study with normal subjects [Ruh *et al.*, submitted] suggest a combination of the two existing computational accounts. Subjects had to learn to perform the routine task of making a cup of coffee or tea in a simulated computer desktop environment. The required objects had to be manipulated on screen with a standard computer mouse by drag and drop, and by single and double clicks. Participants had to discover the order of steps required to make tea or coffee, subject to constraints imposed by the environment, the instructions and their previous knowledge. 40 subjects were tested in two sessions of approximately one hour. The first session was aimed at getting familiar with the virtual environment and learning valid task representation. The data reported here are taken from the second session only. They are assumed to reflect, at least to a certain degree, routinised performance of the task.

The task was held as closely as possible to the task employed in both of the above computational simulations. Task sequences were constructed by concatenating a subset of six invariant sub sequences:

add coffee grounds (7 steps); add teabag (6 steps); add milk (7 steps); add sugar from pack (7 steps); add sugar from bowl (8 steps); drink (4 steps)

Coffee always required adding both milk and sugar, whereas tea was always to be made with sugar only. This leads to four valid coffee sequences:

c1: grounds – sugar from bowl – milk – drink (26 steps)

c2: grounds – milk – sugar from bowl – drink (26 steps)

c3: grounds – sugar from pack – milk – drink (25 steps)

c4: grounds – milk – sugar from pack – drink (25 steps)

and two variations in making tea:

t1: teabag – sugar from pack – drink (17 steps)

t2: teabag – sugar from bowl – drink (18 steps)

In addition to preparing a beverage on screen, subjects had to perform a secondary task – counting how often a certain sound event occurred – in half of the trials. The secondary task served to load attentional resources that are hypothesized to interfere with the function of a supervisory system.

The dependent measure of interest was the response latency at each step of each task, under each condition. Latencies at branching points, i.e., at the first action of a new sub sequence, were of specific interest because the action control system or systems have to determine this step by taking into account (a) the context of task sequence (tea or coffee), (b) the history of getting there (sugar already added or not) and (c) the possible choice of valid sub sequences to enter at this point. Latencies at branching points were therefore compared with those at structurally similar actions within a sub sequence.

The experiment yielded two main results (see fig. 1). Firstly processing times were generally higher at branching points as compared to non-branching points. This supports the hypothesized particularity of these steps in a task sequence. Importantly, though, the size of the effect was dependent on the particular sequence subjects were about to enter. The difference in latencies between branching points and non-branching points depended on task preference. Thus, the difference was highest for the disfavoured case of adding sugar from the bowl, and almost nonexistent for the preferred case of adding sugar from the pack.

Second, we found an interaction of step type and secondary task. Performing the secondary task at the same time prolonged latencies at branching points, but not at non-branching points (see fig. 1). This interaction supports the two systems view. An additional process dependent on attentional resources seems to be involved in processing branching point, even in the preferred task sequences. The distributed representations in an SRN model, on the other hand, provide a more natural account of the fact that latencies at branching points seem to be influenced by preferences and/or familiarity.

3 Combining the approaches

The method of reinforcement learning provides a potential solution to the difficulties with the SRN model identified above. Within reinforcement learning, learning is driven by reward obtained when a specified state is achieved. Reinforcement learning may be extended to sequence learning by also using the prediction of a reward to drive learning. This can be interpreted as implementing goal directed learn-

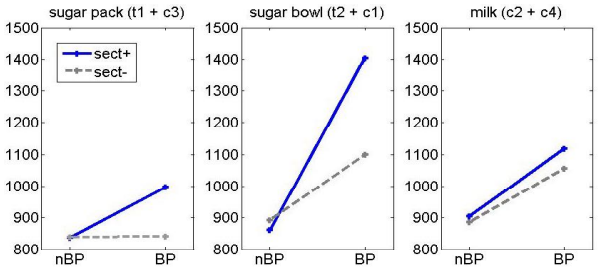


Figure 1: interaction of branching point and secondary task ($F_{pack}(1,78) = 4.055$, $p = 0.047$; $F_{bowl}(1,62) = 8.175$, $p = 0.006$; $F_{milk}(1,55) = 0.158$, n.s.). The non-significance of the last interaction is partly explained by the sparse data in this condition. When subjects who contributed only two examples or less are excluded, the interaction is more evident, though not quite statistically significant ($F(1,20) = 3.581$, $p = 0.073$) due to small sample size.

ing, because the model ultimately learns to approach the rewarded (goal) state while disregarding the means to get there. The same feature also could provide an interface with a supervisory system, as processing difficulties should be indicated by conflicting predictions as to which goals can be reached from a certain state. Enforcing one of the options at this point is exactly the task the hypothesized SAS serves.

By implementing a reinforcement model as an embedded actor/critic architecture with connectionist neural networks, it is possible to preserve the valuable features of the connectionist approach, namely the emergent distributed representations that allow for information sharing, generalisation, context sensitivity, etc. In the remainder of this paper we present a tentative exploration of such a model, showing that it can handle the complexity of a routine task.

3.1 Task

The Nutella task is a simplified version of the coffee/tea task discussed above. Two task sequences must be learned:
Nutellatoast: fixate knife – pick knife – fixate nutella – pick nutella – fixate toast – use knife – pick nutellatoast – eat nutellatoast (8 steps)
Butternutellatoast: fixate knife – pick knife – fixate butter – pick butter – fixate toast – use knife - fixate nutella – pick nutella – fixate buttertoast – use knife – pick butternutellatoast – eat butternutellatoast (12 steps)

3.2 Architecture

An actor/critic architecture was employed with both components implemented as neural networks. The actor was implemented as an SRN with eight units in the input and the output layer, and seven units in the hidden and context layer. Four input units represented whether each of the four available objects (toast, butter, Nutella and the knife) was fixated and the remaining four represented whether the objects were held. The eight output units represented the possible actions (pick, put, use, eat, fixate toast, fixate butter, fixate Nutella, fixate knife).

The critic was implemented as a multi-layer feed forward network with the same 8 input units as the actor, 5 hidden units and one output unit representing the value of the perceived state. A sigmoidal activation function was employed in all layers of both networks.

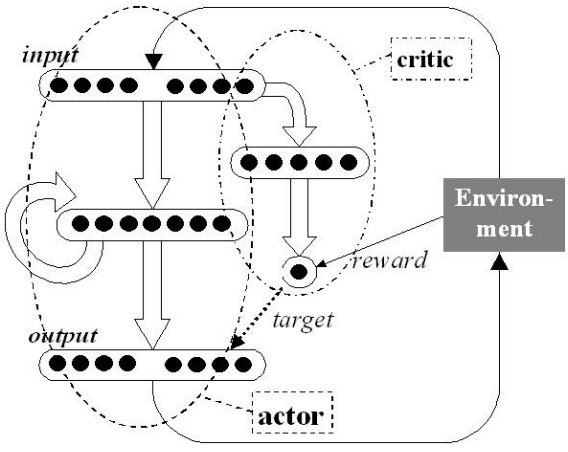


Figure 2: Architecture of the reinforcement model

The actor net was embedded in an environment that maps the chosen actions to the perceived changes in the environment, that is, the new input. If the actor activated the "fixate toast" unit, for example, the unit representing fixation on toast would be turned on in the next input. If an ingredient has already been added to the toast, subsequent fixating on toast would lead to perceiving the toast and the ingredient. The environment also supplied the reward signal. This may be interpreted as being provided by some other part of the cognitive system and not directly by the outside world.

3.3 Learning algorithm

The critic learned to predict the value of the state determined by the chosen action, that is, it approached either the reward it received in the terminal case, or its own prediction at the next step (so-called temporal difference, or TD, learning). By this mechanism, the anticipation of a reward at the end of a sequence is propagated backwards in the sequence. The difference between the prediction and the actual next value is used as an error signal in order to change the weights via backpropagation.

The actor, on the other hand, learned to adjust the activation of the unit that represents the action chosen towards the value predicted by the critic. Only the weights that contributed to the activation of this particular output unit were changed by calculating the difference and propagating this error back through the net.

Both nets learnt at the same time and online, the actor thus "chasing a moving target". In addition, the learning rate was decreased linearly from 0.8 to 0.0 for both nets. The extremely high value at the start is needed for two reasons. Firstly, the nets must make the most out of the rare positive feedback while operating on a very imperfect policy (random behaviour at the start). Secondly, the moving target provided by the critic will be very low, initially, so that a large step towards this value is not a big change in the actor's weight matrix. The learning rate was decreased to allow for fine-grained adjustments towards the end of the learning process.

3.4 Learning regime

Both nets were initialised with small random weights (± 0.5) and none of the input units active. Activation was then propagated through both nets. Uniformly distributed random noise in the range [0.0, 0.5] was added to the activation of each of the actor's output units so as to implement a certain amount of randomness in the network's behaviour. This enables the model to explore the state space. The unit with the highest activation was chosen as the action executed, the new input was obtained via the environmental loop and the prediction of the critic calculated. At this point, the TD-error was calculated and used to adjust the weights of both nets, as described above. The next iteration was then started.

The model starts out with random behaviour, obtaining reinforcement only when it produces a valid (sub) sequence by chance. As the expectation of reward is propagated backwards in time and the actor learns to choose promising actions more often, the critic sees more and more examples of valid sequences and thus is able to establish an ever more accurate estimate of the value of an action, which in turn enables the actor to perform better, and so on.

The model learned on a continuous stream of self produced actions, mediated by the environmental loop. If the actor chose an action that was physically impossible to accomplish (e.g. picking up an object that was already held) the target value for this unit was set to 0, the error propagated back and the choice was repeated. After completion of a task, the context layer was reinitialised.

3.5 Results

Learning

Training consisted of 100,000 iterations of the learning algorithm. Within this period, the model typically produced between 700 and 1100 correct nutellatoast sequences and between 80 and 350 butternutellatoast sequences. The proportion of correct sequences increased steadily until a point of saturation was reached. Similarly, the frequency of negative rewards decreased steadily until it plateaued.

In the trained model, variations occurred in the order of picking up objects, e.g. nutella before knife. This is because the reward only depends on the final state (e.g. eat a nutellatoast). If there are several ways to reach this state, all get rewarded and thus learned. Similarly, the model discovered that in most cases it is more efficient not to put down the knife between task sequences. We return to this important point in the discussion.

Performance

The performance of the model was tested by running 100 iterations using the final set of weights and different levels of noise in the output layer. Without noise, the model typically produced between 12 and 16 correct sequences per 100 iterations (see fig. 3). Minor deviations were observed, usually due to superfluous actions like fixating something else before fixating the required object (e.g. step 10 in fig. 3). Since it is the only non-deterministic feature in this mode of testing, the occurrence of these disturbances must be attributed to the influence of the context layer, which is reinitialised after completion of a task. The fact that the model always finds its way back to the correct behaviour shows its ability not only to produce one fixed sequence, but also to recover quite flexibly from disturbances. Because of the way it is trained, the model has learned more than just one task sequence. In fact, it has acquired knowledge on many different ways that lead to the final aim of receiving a reward, and these different ways include not only variations in the order of actions, but occasional wrong choices as well.

This is even more evident when testing with a higher level of noise. Up to $T \approx 0.3$, structured goal approaching behaviour was produced, although it took increasingly longer for the model to find its way back to completing a certain sequence. For even higher levels of noise, the model only occasionally found its way into a sequence attractor's basin. It is interesting to note that disturbances occur more often at or before the beginning of a sequence than towards the end. This is because values of states early in a sequence are low, and the corresponding action units thus cannot gain a large advantage over the other output units. Secondly, the influence of the randomly initialised context layer is bigger shortly after resetting. Furthermore, the choices of actions that lead to a certain outcome are more tightly constrained at later stages. One might start in different ways to make nutellatoast, but picking it up and eating it will always be the unique possible end. This feature of the model's behaviour leads to testable predictions on human action sequencing. Specifically it suggests that humans are more likely to commit errors near the beginning rather than the end of a sequence and that action selection will get faster towards the end of a task sequence.

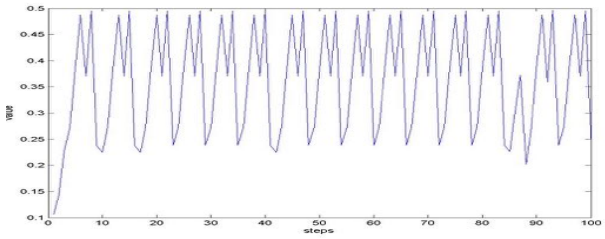


Figure 3: The value function of the critic when tested without noise. The regular patterns with two peaks correspond to the nutellatoast sequence. The three-peak pattern around step 90 is an example of the butternutellatoast sequence.

Representations

Visualization of the trajectory in activation space of the actor by means of multi dimensional scaling (fig. 4) shows the "shadowing" typical for recurrent networks (see Botvinick & Plaut, 2004). The last six steps of the two sequences in figure 4 resemble each other closely, despite the fact that the Nutella is added to the plain toast in one case, but to the toast with butter already on it in the other case (i.e. different input). Also, the adding-butter subsequence (steps 1–4 in the dotted trajectory) is similar to adding Nutella (steps 5–8). This picture exemplifies the ability of the network to encode structural similarity in sub sequences and to share information between them.

Key Parameters

While the model has several parameters, two are of particular importance [see Ruh *et al*., to appear, for more details]:

Noise (T): Within all reinforcement learning models there is a trade-off between exploitation of acquired knowledge (the "policy") and exploration of uncharted areas of the state space. Exploration is needed to improve the policy, but it leads to unnecessary mistakes in cases where behaviour is already adequate. In the present model the trade-off was implemented by adding random noise, uniformly distributed between 0 and T, to the activations of output units prior to

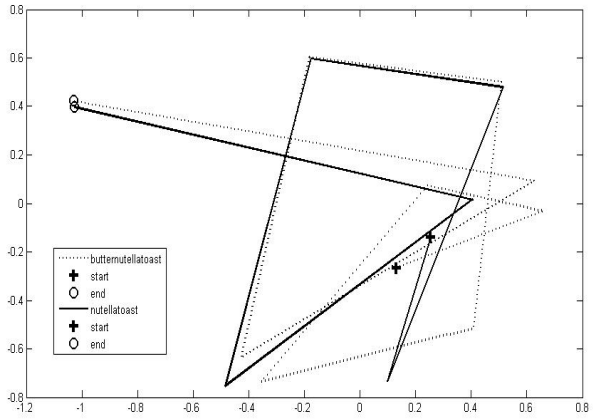


Figure 4: Multi-dimensional scaling plot of two task sequences performed by the model when tested without noise.

selecting the most active output unit as the next action. This implementation is equivalent to the temperature parameter T in a probabilistic softmax function. The choice is more random if there are many equally highly activated units, and not random at all if only one unit is maximally excited but none of the others are. With T held constant, this still leads to an increasingly structured behaviour of the actor as its output activations become more differentiated.

Decreasing T during learning leads to more correct sequences (up to 10 times more) in the learning phase, but not to better performance when tested without noise. In fact, these nets were easier to divert, presumably because they lacked experience in recovering from wrong choices.

Note that in the standard case with T = 0.5, the model's perfect behaviour when tested without noise arose out of an overall rather poor performance during learning. This so-called off-policy learning shows the relative independence of critic and actor and has interesting implications for a psychological view of learning. Specifically it suggests that there is much to be learned from doing things wrongly.

Negative reward: When the actor attempted to perform a physically impossible action (e.g. picking up when nothing is fixated or too many objects are already held) the activation of the chosen output unit was adjusted towards zero (i.e., it received a negative reward). A naïve view would be that negative rewards, if used efficiently, should lead to a substantial restriction of the state space that the model has to explore. This was not the case in the present simulation, though, as values and activations of all actions are driven to near zero during the first few hundred iterations and rise only slowly afterwards. Hence impossible actions keep on getting chosen. The essential role of the negative reinforcement term (without it learning was unreliable – some models did not learn to perform the task at all) proved to be more complex: When, at the beginning of learning, a reward is encountered, the weights in the actor that led to the chosen action are changed to a large extent because of the high learning rate. This inevitably leads to the situation that different inputs inappropriately will also elicit a higher activation of that same action. This ability to generalize can be desirable if applied to "similar" situations, but because of the net's initial inability to correctly categorize situations in combination with the high learning rate, generalization will be applied too widely. In the present model, valid experience is a valuable thing, because it is dependent on the current behaviour of the actor. Furthermore, experience is mainly based on positive evidence (i.e., reward given or expected). In this learning regime it is very hard to discover that an overly valued state really is not so good after all. By introducing the negative reward term, the actor is given a new, more direct and more frequent source of experience that helps to correct the initial overgeneralization. Owing to this mechanism the net can take more information out of the occasional positive reward signal and at the same time is able to exhibit better behaviour and thus to provide more positive evidence.

4 Discussion

The present routine task, by being naturalistic, involves many complexities beyond the standard theoretical tasks frequently considered in the reinforcement learning literature: a large state space (~ 8^8 states); many possible actions/output states; non-Markovian states (e.g. if butter and toast are held, is the butter on the toast or were they picked up separately?); and relatively long sequences and thus long distance dependencies. The simulations reported here show that it is possible to learn this structurally quite complex task within a relatively simple model with few *a priori* assumptions in the comparatively short time of 100,000 iterations. The power of this model arises from the interplay of different principles, each of which simplifies the learning of the task at hand in its own fashion:

- Because the model is embedded in an environment, some of the possible states in state spaces are not accessible and thus do not have to be taken into account.
- Because of the recurrent connections in the actor, information of earlier states can be preserved and therefore can turn the decision at a later state into a Markov Decision Problem.
- The actor can use the knowledge acquired by another part of the model, the critic, as positive evidence that helps it to improve performance and thus leads to more correct examples to learn from for the critic. The relative independence of the two nets guarantees good learning even for imperfect policies.
- The use of negative evidence for one-step sequences can furthermore improve the actor's behaviour.
- Reward for shorter sequences can bootstrap the model to the acquisition of longer sequences. Generalization, for valid, as well as for invalid actions, helps the model to learn different sequences with partial overlap more efficiently.

The task representations the actor develops are comparable in many respects to the ones in an SRN that uses supervised learning: they overlap structurally (generalization), they allow for information sharing between similar (sub) se-

quences, they are graded and context sensitive etc. However, in contrast to a standard SRN, a reinforcement model actively explores the whole state space and thus is able to recover flexibly from wrong choices along the way. As an additional advantage of this mechanism one might expect such a model to be less prone to catastrophic interference, because a certain amount of experience with all physically possible sequences is implicit in the final pattern of connection weights.

The model is not dependent on a carefully balanced training set that provides it with an equal amount of experience for every task version it is to produce because the possible variations in a task are discovered by trial and error. Rather than forming attractors for the valid examples and organizing the remainder of the state space in terms of proximity to them, a reinforcement model experiments with many different sequences and discovers which choices to make in order to get rewarded in the end. Even if the resulting representations resemble each other in many ways, in a reinforcement model they develop because it tries to reach rewarded states (i.e., goals), not because it adapts to the examples it is taught with.

Evidently, more work needs to be done before the present model can give a full-scale account of hierarchical routine action. Several adjustments could help the model to learn more efficiently, such as bootstrapping the learning process with a few valid examples of task sequences, or adjusting the learning rate and/or the exploration/exploitation ratio dynamically according to some measure of current performance. Botvinick *et al.* [2001] propose a measure of conflict that seems suitable to indicate how well the model is doing at a certain point in time. High conflict indicates inefficient behaviour and therefore might trigger an increased learning rate, while low conflict would indicate that the policy is adequate and should not be changed.

Another issue seems to be more important though: so far, all the rewards given simply are numerical values along one dimension. If one sees the value maximizing behaviour of the nets as corresponding to goal directed behaviour in humans as suggested by the reinforcement literature, then, so far, there is only one single goal. The critic is unable to discriminate between the rewarded sequences, because they all influence its single output unit. The aim of our ongoing work is to implement multiple goals, as well as a way to dynamically switch between them. In a recent model of the Wisconsin card sorting task [Rougier & O'Reilly, 2002], the adaptive critic (AC) component plays this role by influencing the 'sorting rule' the network applies via lateral connections to the hidden layer. Transferred to our model this means a way to tell the net which one of several value functions, each representing a different goal, is to be maximized at a certain point in time. Importantly, this would result in one goal per task, not one 'instruction' per task version.

In summary we have shown that a reinforcement approach to modelling the basic system involved in hierarchical routine action selection is promising. The presented model shares the advantages of the emergent representations in supervised connectionist architectures, but has additional benefits in terms of plausibility of the process of learning, flexibility of representations and the ability to account for goal directed behaviour. The next step is to develop an activation based supervisory component, similar to Rougier & O'Reilly's [2002] AC/PFC module that is able to enforce the pursuit of a certain goal by actively nudging/biasing the basic system's hidden activations into the influence of a certain attractor's basin. Importantly, the information of which goal can be achieved from the current state is readily available in the system in terms of the value function(s), thus potentially providing an interface between the basic system and a supervisory component.

References

[Botvinick and Plaut, 2004] Matthew M. Botvinick and David C. Plaut. Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111:* 395–429, 2004.

[Botvinick and Plaut, 2002] Matthew M. Botvinick and David C. Plaut. Representing task context: proposal based on a connectionist model of action. *Psychological Research*. 66: 298–311, 2002.

[Botvinick *et al.*, 2001] Matthew M. Botvinick, Todd S. Braver, D.M. Barch, C.S. Carter, and J.D. Cohen. Conflict Monitoring and Cognitive Control. *Psychological Review*. 108(3): 624–652, 2001.

[Cooper and Shallice, 2000] Richard P. Cooper and Tim Shallice. Contention Scheduling and the control of routine activities. *Cognitive Neuropsychology, 17*: 297–338, 2000.

[Norman and Shallice, 1986] Donald A. Norman and Tim Shallice. Attention to action: willed and automatic control of behaviour. In R. Davidson, G. Schwartz, and D. Shapiro (eds.) Consciousness and Self Regulation, Volume 4. Plenum: NY, 1986.

[Reason, 1984] James T. Reason. Lapses of attention in everyday life. In Parasuraman, W., & Davies, R. (ed.), *Varieties of Attention*, ch. 14: 515–549. Academic Press, Orlando, FL, 1984.

[Rougier and O'Reilly, 2002] Nicolas P. Rougier and Randall C. O'Reilly. Learning representations in a gated prefrontal cortex model of dynamic task switching. *Cognitive Science*. 26: 503–520, 2002.

[Ruh *et al.*, to appear] Nicolas Ruh, Richard P. Cooper and Denis Mareschal. The time course of routine action. Submitted to the 27th Annual Conference of the Cognitive Science Society. Italy, July 2005.

[Ruh *et al.*, to appear] Nicolas Ruh, Richard P. Cooper and Denis Mareschal. A Reinforcement model of sequential routine action. To appear in *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*. Finland, June 2005.

Modeling routine sequential action with recurrent neural nets

Matthew M. Botvinick
University of Pennsylvania
Department of Psychiatry and Center for Cognitive Neuroscience
3720 Walnut St., Philadelphia, PA 19104
mmb@mail.med.upenn.edu

Abstract

The performance of everyday sequential tasks presents deep computational challenges which have been acknowledged both by computer scientists and by psychologists and neuroscientists. Since the 1950's a number of computational accounts have been put forth to account for routine sequential action, attending to varying degrees to the issue of neural implementation. In general, such accounts have involved hierarchies of schemas or processing units, mirroring the hierarchical structure of everyday tasks themselves. In recent work, we have explored an alternative approach, according to which hierarchically structured behavior emerges from a recurrent network architecture, mapping from perceptual inputs to action outputs. Implementations of the theory demonstrate that it can account for numerous central aspects of human sequential action, including patterns of error in both normal and apraxic performance. A key finding is that hierarchically structured behavior can emerge from a processing system that is not itself hierarchically structured. Having established this, we explored, in further simulations, the consequences of introducing architectural hierarchy into the framework. The results of these simulations point toward a novel hypothesis concerning the development of prefrontal cortex, linking its role in temporal integration to its position within a hierarchy of cortical areas.

1 Introduction

Washing the dishes, making the bed, getting dressed to go outside in the rain: Familiar routines such as these occupy a great deal of daily life. Because such action sequences are typically negotiated with relative ease, one tends to overlook their underlying psychological complexity. In fact, everyday naturalistic behavior requires the highly coordinated use of hard-won perceptual and motor skills, semantic memory, working memory, and attentional control. Perhaps it is not surprising, given such complexity, that the neural mechanisms underlying routine sequential behavior remain only partially understood.

A key challenge in everyday sequential behavior, which has been the focus of much theoretical discussion, is the fact that many sequential routines display a roughly hierarchical structure, being composed of low-level actions organized into goal-directed subtasks, which are in turn organized into larger overall tasks (Figure 1). A key question is how the brain manages to deal with such hierarchical structure.

In recent work, my colleagues and I have proposed a computational account of routine sequential action that seeks to address this basic problem in terms that can also be mapped, in a general way, onto neural circuitry. According to this framework, routine sequential action arises from a massively parallel neural system that maps from perceptual inputs to action outputs via learned, distributed internal representations. This system, it is assumed, is further characterized by extensive recurrent connectivity, which allows information about temporal and task context to be preserved over time.

In what follows, we summarize the results of computer simulations and empirical work, evaluating the ability of this framework to account for fundamental

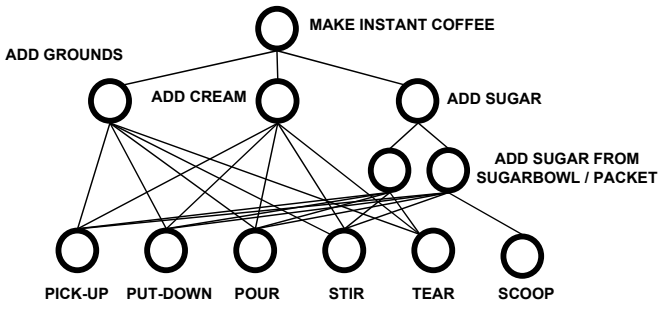


Figure 1. A hierarchical model of the task of making coffee. From Cooper and Shallice (2000).

properties of human behavior in routine sequential domains.

2 A model of routine sequential action

Botvinick and Plaut [2004] report a set of simulations in which our basic theory was implemented in the form of a recurrent connectionist network, applied to a set of specific everyday tasks. Our interest in this study was in whether a recurrent network without explicit hierarchical structure could handle hierarchically structured sequential tasks. This question was of particular interest, given that previous accounts of sequential action had traditionally assumed that the processing system itself must assume a hierarchical form, with discrete elements coding for entire task and subtask sequences [see Figure 1; Cooper and Shallice, 2000]. An additional set of questions related to errors in routine sequential behavior. In particular, it was asked whether the model could account for key properties of everyday slips of action, and for patterns of error seen in patients with action disorganization syndrome, a type of apraxia affecting performance in sequential routines [Schwartz et al., 1998].

2.1 Model architecture and task domain

The structure of the model is diagramed in Figure 2. Like all connectionist-style neural network models, it is comprised of simple processing units, each with a scalar activation value. These excite or inhibit one another through adjustable, weighted connections. In the current model, units are organized into three groups. A group of input units serves to represent the perceptual features of objects in the environment. These units connect to an internal or 'hidden' group, which itself connects to an output group whose units represent simple actions (e.g., 'pick-up,' 'pour', or 'locate-spoon'). In order to capture the fact that actions affect perceptual inputs, the model communicates with a simulated environment, which updates inputs to the network contingent on selected actions.

A crucial feature of the model is that there are reciprocal connections between each pair of units in its internal layer. The presence of these 'recurrent' connections means that activation can flow over circuits within the network, allowing information to be preserved and transformed over multiple steps of processing. It also has important implications for the role of the model's internal units. Given their overall pattern of connectivity, these units play two roles. First, they serve as an intermediate stage in the stimulus-response mapping performed on each processing step. Second,

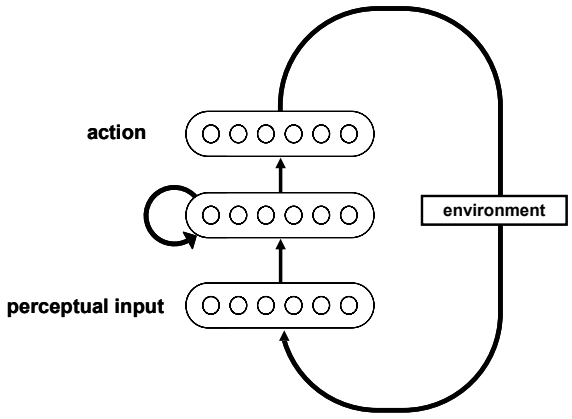


Figure 2. Architecture of the model used by Botvinick and Plaut (2004). Arrows indicate all-to-all connections. The input layer contains 39 input units, each coding for an object descriptor. Multiple units are activated in this layer to describe the currently viewed and held objects (e.g., "packet," "paper" and "torn"). The output layer contains 19 units, each representing an action (e.g., "pour" or "fixate-spoon"). The hidden layer contained 50 units.

because they carry all of the information that will be conducted over the network's recurrent connections — and thus all of the information that will be carried over to the next time-step — they are responsible for carrying the model's representation of temporal context.

A number of studies have demonstrated the ability of recurrent networks to address aspects of human behavior in the domains of language [e.g., Elman, 1990] and implicit learning [e.g., Cleeremans, 1993]. Our simulations investigated whether similar computational principles could be used to account for human behavior in everyday, goal-oriented tasks involving the manipulation of objects. In order to facilitate comparison with the recent hierarchical model of Cooper and Shallice (2000), the task modeled was that of making a cup of instant coffee. Our implementation of the task is shown schematically in Figure 3. It comprises four subtasks, each containing between five and eleven actions: (1) adding coffee-grounds, (2) adding cream, (3) adding sugar (by one of two methods), and (4) drinking. For reasons that will become clear in later discussion, the training corpus also contained a second task, tea-making. The model was trained to perform these tasks using a version of the backpropagation learning algorithm [Williams and Zipser, 1995]. Training was analogous to observing and attempting to predict the sequence of actions of a skilled individual repeatedly carrying out specific versions of each task. Testing involved successively presenting the trained model with perceptual input and using its generated action to

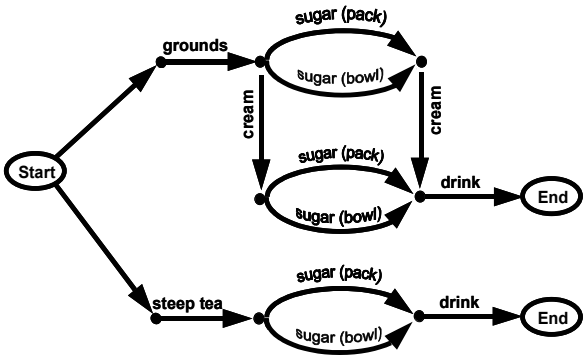


Figure 3. Structure of the coffee and tea tasks. Arrow-segments represent sequences of between 5 and 11 actions.

modify the environment (and, hence, the model's subsequent perceptual input).

2.2 Overview of simulation results

In our simulations of normal performance, we asked simply whether the model could learn to perform the target tasks. Some action researchers have expressed doubt concerning the ability of recurrent networks to deal with tasks that are hierarchically structured, that is, tasks made up of subtasks and actions that also appear as part of other tasks [see, e.g., Houghton and Hartley, 1995]. Consistent with earlier studies applying recurrent networks in hierarchical domains, the model proved quite capable of learning the target sequences, and producing them autonomously following training. Our simulations of action slips and ADS were based on the assumption that both stem from disruptions to representations of temporal or task context. In our model, as noted above, such context information is carried by the hidden units. With this in mind, context information was degraded by randomly perturbing the activation values in the hidden layer on each cycle of processing. When this was done mildly, the model produced errors resembling human slips of action. In line with empirical observations concerning slips [Norman, 1981; Reason, 1990], the model made errors at decision points, behavioral 'forks in the road' where the actions just completed bear associations with multiple lines of subsequent behavior. Also like typical human slips, the model's errors took the form of subtask sequences performed correctly but in the wrong context. The model's errors fell into the same categories as human slips: omissions, repetitions, and lapses from one task into another. With increasingly severe disruption to the model's context representations, the model's behavior became gradually more fragmented, coming to resemble the performance of ADS patients as characterized in recent empirical studies [e.g., Humphreys *et al.*, 2000; Schwartz *et al.*, 1998].

Botvinick and Plaut (2004) point to a number of apparent advantages of the model over traditional accounts of routine sequential action. Some of these pertain to the model's ability to capture particular behavioral phenomena. Specifically, the model produced at least one type of error (recurrent perseveration) not observed in the simulations of Cooper and Shallice (2000); it reproduced a correlation between error rate and the distribution of error types reported by Schwartz et al. (1998), another effect not captured by Cooper and Shallice (2000); and, again unlike that earlier study, the Botvinick and Plaut (2004) model displayed a smooth variation in behavioral fragmentation with damage, a feature of ADS. Botvinick and Plaut (2004) also discuss several other advantages of the model over traditional accounts, including its reliance on learning instead of extensive 'hand wiring,' its avoidance of the inflexible, ad hoc sequencing mechanisms typically incorporated into traditional models, and its relative strength in dealing with context-sensitive behavior.

In order to understand how the model works, and why it makes the errors that it does, it is necessary to consider how the model represents task context within its internal or hidden layer. We turn now to a discussion of this issue.

2.3 Representations of task context

Whether the model is used to simulate normal performance or errors, its behavior is linked directly to the patterns of activation over the units in its internal layer. As noted above, these units play two roles. Because they lie between input and output layers, they are responsible for facilitating the stimulus-response mapping being performed on each time-step. Second, because — via their recurrent connections — the internal units transmit information from one time-step to the next, they also must serve to represent the current behavioral context. In this sense, the patterns of activation arising in the model's internal layer play the role that is played, in traditional models, by task and subtask nodes; on each time-step, the information carried in this layer is integrated with information about external inputs in order to determine the context-appropriate action. Note that every unit in the hidden layer participates in each context representation. Unlike hierarchical models of action, which use single units to represent entire task contexts, the present model employs distributed representations [see Hinton

et al., 1986]; information is represented by an entire population of processing units, within which each unit participates in representing a variety of contexts.

In order to understand the implications of the model's way of representing context, it is useful to adopt a spatial metaphor. The model's internal layer contains 50 units, each of which carries an activation between zero and one. If these activations are thought of as spatial coordinates, then each pattern of activation (context representation) can be thought of as specifying a point in a 50-dimensional representational space. As the model steps through an action sequence, the successive patterns in its internal layer can be thought of as tracing out a trajectory in this space. Although it is impossible to visualize such trajectories in their original 50 dimensions, one can gain a sense of them using the technique of multi-dimensional scaling (MDS) [see Kruskal and Wish, 1978]. An example of the model's internal representations, visualized in this way, is shown in Figure 4. The plot shows two trajectories, both representing the sequence of internal states produced by the model as it stepped through the eleven actions of the sugar-adding subtask, in one case during coffee-making, and in the other during tea-making. The first thing to note is that the two trajectories are similar in shape. This indicates that the series of internal representations the model uses when adding sugar to coffee are similar to those it uses when adding sugar to tea, an arrangement that makes sense since sugar-adding involves the same sequence of actions regardless of the overall task context. However, the two trajectories are not precisely identical. The minor differences between the two reflect the difference in overall task context; the model's internal representations on each step differ slightly according to whether it is coffee- or tea-making that is being performed. As earlier studies of recurrent networks [e.g., Servan-Schreiber *et al.*, 1991] have expressed it, the network "shades" its internal representations to reflect differences in context. It is in this way that the model manages to maintain important information about temporal context, while at the same time dealing with immediate stimulus-response mappings.

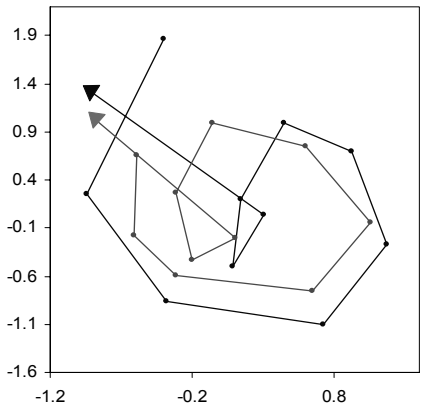


Figure 4. Multidimensional scaling analysis of the internal representations arising during performance of the sugarpack subtask, either in the context of coffee-making (black) or tea-making (gray).

2.4 Testing a prediction of the model

As noted above, errors in the Botvinick and Plaut model — as in human performance — tend to fall at decision points, i.e., the transitions between subtasks. The reason for this can be understood in terms of the representational 'shading' just discussed. Transitions between subtasks typically require accessing information about the larger task context. For example, in the task of coffee-making, when one completes adding cream, knowing what to do next depends on knowing whether sugar has yet been added. In the model, such information is preserved through the shading of representations in the hidden layer. However, because the model's hidden units are subject to a small amount of noise, context information can be lost before it is needed, resulting in an error.

Importantly, according to the model, the degradation that leads to decision-point errors can occur at any time, not only at the boundaries between subtasks. Indeed, a distinctive claim of the Botvinick and Plaut model is that context information is most susceptible to loss toward the middle of subtask sequences. The explanation for this relates to how differences in temporal or task context are internally represented. In the model, distinctions between different contexts are represented very robustly close to decision points, where such distinctions are directly relevant to action selection. However, elsewhere, and in particular toward the middle of subtask sequences, differences in temporal context are represented less strongly. This, in turn, makes it easier for the system's representation of context to become disrupted, setting the scene for a later decision-point error.

This aspect of the Botvinick and Plaut theory leads it to make a distinctive prediction about the effect of momentary distraction. Specifically, the model predicts that distraction should be most disruptive when it falls toward the middle of a subtask sequence, even though the errors that result from such distraction do not occur until the end of the subtask. This prediction is of some interest, because earlier theories of action slips would appear to lead to the opposite prediction. Previous accounts of decision-point errors have attributed them to the failure of specific memory-retrieval operations occurring at the decision point itself [Norman, 1981; Reason, 1990]. Such theories would presumably predict that distraction would be more disruptive when it falls near a decision point than when it falls further away.

In order to test the predictions of the Botvinick and Plaut theory, Botvinick and Bylsma [in press] studied the performance of normal participants on an everyday task (coffee-making), under conditions involving intermittent distraction. Distraction was imposed by momentarily interrupting subjects' performance on this task, requiring them to perform a secondary arithmetic task. Interruptions were timed to occur either at the end of a subtask (i.e., just before a decision point) or at mid-subtask. As predicted based on our computational work, the frequency of action slips, and, in particular, decision-point errors, depended on where the distraction occurred relative to subtask boundaries. A greater number of decision-point errors followed interruptions occurring midway through a subtask than following interruptions falling at the end of a subtask.

3 Addressing a neuroanatomic division of labor

3.1 Fuster's hierarchy

A distinctive aspect of the Botvinick and Plaut (2004) model is that information concerning task context is represented over the same population of units that mediates immediate input-output mappings. The model thus demonstrates the point that performance in hierarchically structured domains does not require a hierarchically structured processing system, as has been assumed in most psychological work on routine sequential behavior. However, it is of course a separate question whether this point is relevant to the brain. Neuroscience clearly does provide some empirical evidence to support the idea that context information is represented at the same level as more immediate representations of action. For example, Aldridge and Berrige [1998] showed that neurons in rat striatum show different patterns of activity during grooming movements, depending on whether those movements occur in isolation or as part of a larger grooming sequence. Nevertheless, there is also considerable evidence for some degree of functional segregation in the brain. In particular, the prefrontal cortex (PFC) seems to play a special role in representing task context information. This role has been stressed in the context of naturalistic action by Grafman and colleagues [Grafman, 2002; Zalla *et al.*, 2003]. On a more general level, Fuster [1997] has characterized the representation of temporal context information as a defining function of the PFC. Interestingly, Fuster also portrays the PFC as occupying the apex of a hierarchy of cortical regions. At the base of this hierarchy are primary sensory and motor areas, concerned predominantly with representing immediate inputs and outputs. Above this are secondary sensory and motor areas; above this association cortices; and above this the PFC (Figure 5, top).

In modeling hierarchically structured behavior, Botvinick and Plaut [2004] eschewed architectural hierarchy. However, an interesting question arises in the light of Fuster's account: What would happen if some degree of architectural hierarchy were built into the model, from the outset? It would be interesting if a model structured in the form of Fuster's hierarchy (Figure 5, top) developed, through learning, a functional division of labor, with units at higher levels playing a relatively more important role in maintaining context information.

In order to evaluate this possibility, we constructed a neural network model containing seven groups of

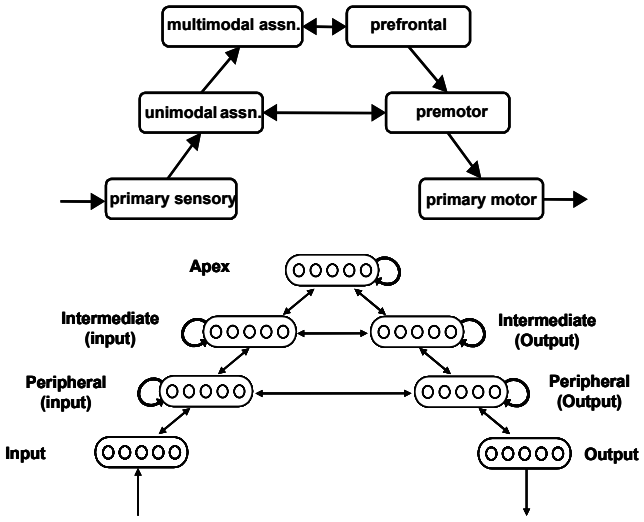


Figure 5. Top: A hierarchy of cortical regions, based on the account of Fuster (1997). Bottom: Architecture of a model based on Fuster's hierarchy.

units, interconnected as shown in Figure 5 (bottom). The model was implemented as a fully recurrent attractor network running in continuous time, and was trained using gradient-descent learning [recurrent backpropagation through time; Williams and Zipser, 1995]. The task was chosen in order to allow a direct assessment of the degree to which units in the model coded for immediate inputs and outputs vs. temporal context information. Specifically, we trained the model on the store-ignore-recall (SIR) task [O'Reilly and Munakata, 2000]. This involves presentation of a extended series of individual digits. If the digit presented is shown in black, the task is simply to read the number aloud. If the digit is shown in red, again the digit is read aloud, but it is also to be held in memory until the appearance — following a variable number of intervening digits — of a recall cue, in response to which the stored number is to be reported. Digits were represented using individual input units. Two additional units were also included, one to indicate "red" and the other representing the recall cue. In training the model, target activations were applied at both input and output. That is, the model was trained not only to produce the correct response, but also to send activation to the input layer consistent with the present input.

Following training, the model showed perfect performance on the SIR task. However, our real interest was not in the model's overt performance, but in the representations underlying it. In particular, we wished to evaluate the degree to which each group of units in the model was involved in representing stored context information, as opposed to representing immediate inputs and outputs. To this end, we recorded the activation of each unit during processing of black digits that occurred between a red digit and the recall cue. We then evaluated the degree to which this activation varied depending on 1) the identity of the black digit, and 2) the identity of the earlier red digit. The ratio of these two, which we refer to as the coding ratio, provided an index of the degree to which units were involved in storing context information. The average coding ratio was computed for each unit group.

As shown in Figure 6 the coding ratio was found to vary rather widely across groups, growing progressively larger with each step up in the architectural hierarchy. Thus, the model developed through learning a regional differentiation of function like the one described by Fuster (1997), with processing structures at higher levels of the hierarchy — and in particular at its apex — coding preferentially for temporal context information. This division of labor was apparent in the behavior of the model, as well. When units in the apical group were lesioned, the coding ratio at lower levels fell, and the model's recall performance deteriorated.

It is important to emphasize that the division of labor illustrated in Figure 6 emerged quite spontaneously as a result of learning. There was nothing in the construction of the model that prevented context information from being handled entirely at lower levels of the hierarchy; indeed, a version of the model that contained only the groups labeled "peripheral" in the figure was found to be entirely capable of acquiring the task. The emergence of the division of labor in the model appears to reflect differing pressures on unit groups during learning, as a function of their synaptic distance from the periphery. The groups directly connected to the input and output layers are immediately responsible for generating the correct pattern of activation in those layers, and are thus under pressure to strongly represent current inputs and outputs. With immediate input-output mappings handled by lower-level groups, groups further from the periphery are freed up to represent context information. Indeed, it makes sense to represent such information away far from the periphery, since there are many steps in the task during which context information is irrelevant to response selection.

Despite its simplicity, this simulation reveals an interesting possibility concerning the relationship between the function of the PFC and its connectivity with other parts of the brain. Fuster (1997) stressed, on the one hand, the involvement of the PFC in representing temporal context, and on the other hand the position of the PFC at the apex of a hierarchy of cortical areas. Our simulation provides a motivation for the hypothesis that these two aspects of PFC are closely interrelated, and in particular that the connec-

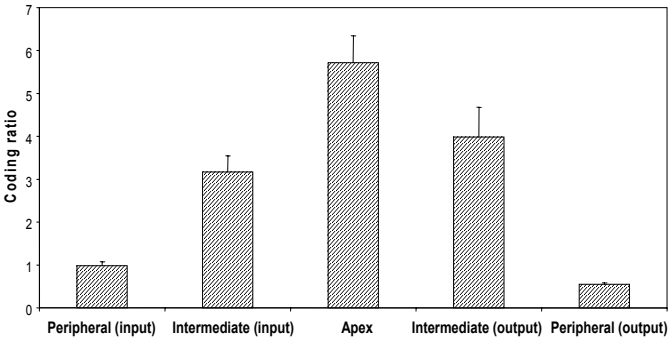


Figure 6. Coding ratios observed in each unit group of the model depicted in Figure 6 (bottom).

tivity of the PFC may provide an explanation for why it comes to assume a role in representing temporal context.

3.2 Application to naturalistic sequential behavior

In a follow-up simulation, we tested the effect of including a minimal architectural hierarchy in the Botvinick & Plaut (2004) model of naturalistic sequential action. To this end, an additional group of units was added to this model, as shown in Figure 7 (top). This group connected only to the original hidden layer, and was thus at a greater synaptic distance from the periphery (input and output layers) than the latter. As in the SIR simulation, the question was whether units further from the periphery, i.e., the units in this new layer, would assume a special role in representing task context. In order to evaluate this, the model was trained on the coffee and tea tasks used by Botvinick and Plaut. Unit activations were then measured in each hidden layer during performance of the sugarpack subtask. Sensitivity to immediate input-output mappings was measured in terms of the change in each hidden layer's pattern of activation with successive steps in the subtask (Figure 7, middle). Sensitivity to task context was measured in terms of the degree to which the pattern of activation on each step of the subtask differed depending on the task context (coffee or tea; Figure 7, bottom). As in the SIR model, the unit group further from the periphery was found to code preferentially for context information.

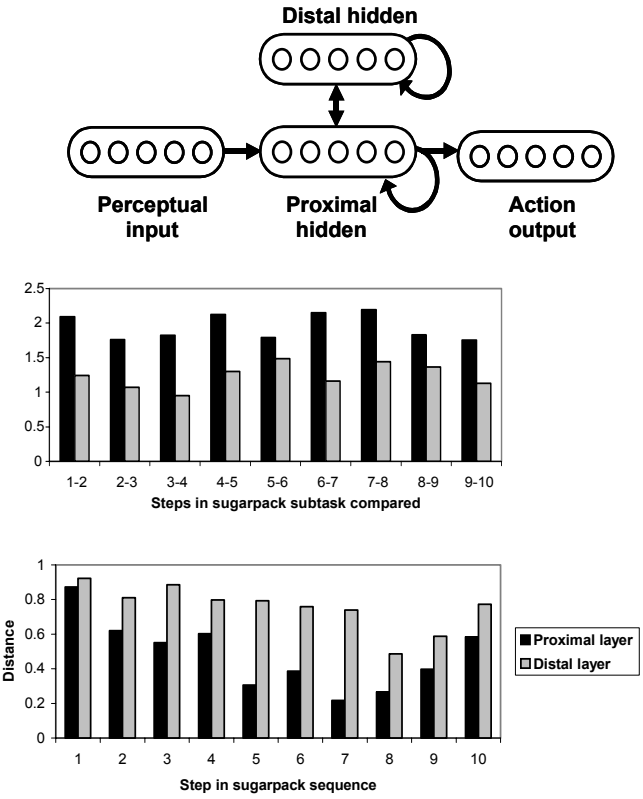


Figure 7. Top: reimplementation of the Botvinick and Plaut (2004) model, with a new hidden layer. Bottom: Cartesian distances between pairs of internal representations.

4 Conclusion

The present paper has summarized a set of simulations, and a bit of empirical work, focused on a particular computational account of routine sequential behavior. According to this account, highly familiar everyday tasks are accomplished by a system that maps from perceptual inputs to motor outputs, via internal representations that integrate and maintain information concerning temporal and task context. The capacity of this processing system to preserve and transform context information over time inheres in massively recurrent connectivity. The initial goals of the modeling project were to demonstrate how the computational properties of such a system might give rise to key aspects of human performance in hierarchically structured naturalistic domains. Given the success of this enterprise, it is of interest to consider how the computational principles involved in the model might relate to the neural structures underlying routine sequential behavior. Clearly, there is an analogy between the massively recurrent connectivity involved in our models and the feedback loops connecting cerebral cortex with basal ganglia and thalamus [Middleton and Strick, 2000], loops that have been proposed to play a critical role in guiding sequential action [Houk and Wise, 1995; Tanji, 2001]. The simulations implementing Fuster's hierarchy suggest how massive recurrence combined with specific patterns of regional connectivity might also contribute to the specific role posited for prefrontal cortex in guiding sequential behavior.

Acknowledgments

The present work was supported by National Institute of Health award MH16804.

References

[Aldrige & Berrige, 1998] W. J. Aldridge, and K. C. Berrige. (1998). Coding of serial order by neostriatal neurons: a "natural action" approach to movement sequence. *Journal of Neuroscience, 18*, 2777-2787.

[Botvinick and Bylsma, in press] M. Botvinick, and L. M. Bylsma. (in press). Distraction and action slips in an everyday task: Evidence for a dynamic representation of task context. *Psychonomic Bulletin and Review*.

[Botvinick and Plaut, 2004] M. Botvinick, and D. C. Plaut. (2004). Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review, 111*(2), 395-429.

[Cleeremans, 1993] A. Cleeremans. (1993). *Mechanisms of implicit learning: connectionist models of sequence processing*. Cambridge, MA: MIT Press.

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

[Elman, 1990] G. Elman. (1990). Finding structure in time. *Cognitive Science, 14*, 179-211.

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

[Grafman, 2002] J. Grafman. (2002). The structured event complex and human prefrontal cortex. In D. T. Stuss & R. T. Knight (Eds.), *Principles of frontal lobe function* (pp. 292-310). London: Oxford University Pressq.

[Hinton *et al.,* 1986] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. (1986). Distributed representations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

[Houghton and Hartley, 1995] G. Houghton, and T. Hartley. (1995). Parallel models of serial behaviour: Lashley revisited. *Psyche, 2*.

[Houk and Wise, 1995] J. C. Houk, and S. P. Wise. (1995). Distributed modular architecture linking basal ganglia, cerebellum and cerebral cortex: its role in planning and controlling action. *Cerebral Cortex, 5*, 95-110.

[Humphreys *et al.,* 2000] G. W. Humphreys, E. M. E. Forde, and D. Francis. (2000). The organization of sequential actions. In S. Monsell & J. Driver (Eds.), *Attention and Performance XVIII* (pp. 425-472). Cambridge, MA: MIT Press.

[Kruskal and Wish, 1978] J. B. Kruskal, and M. Wish. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage Publications.

[Middleton and Strick, 2000] F. A. Middleton, and P. L. Strick. (2000). Basal ganglia and cerebellar loops: motor and cognitive circuits. *Brain Research Reviews, 31*, 236-250.

[Norman, 1981] D. A. Norman. (1981). Categorization of action slips. *Psychological Review, 88*, 1-14.

[O'Reilly and Munakata, 2000] R. C. O'Reilly, and Y. Munakata. (2000). *Computational explorations in cognitive neuroscience: understanding the mind by simulating the brain*. Cambridge: MIT Press.

[Reason, 1990] J. T. Reason. (1990). *Human error*. Cambridge, England: Cambridge University Press.

[Schwartz *et al.,* 1998] M. F. Schwartz, M. W. Montgomery, L. F. Buxbaum, S. S. Lee, T. G. Carew, and H. B. Coslett. (1998). Naturalistic action impairment in closed head injury. *Neuropsychology, 12*, 13-28.

[Servan-Schreiber *et al.*, 199] D. Servan-Schreiber, A. Cleeremans, and J. L. McClelland. (1991). Graded-state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning, 7*, 161-193.

[Tanji, 2001] J. Tanji. (2001). Sequential organiation of multiple movements: involvement of cortical motor areas. *Annual Review of Neuroscience, 24*, 631-651.

[Williams and Zipser, 1995] R. J. Williams, and D. Zipser. (1995). Gradient-based learning algorithms for recurrent neural networks and their computational complexity. In Y. Chauvin & D. E. Rumelhart (Eds.), *Backpropagation: Theory, architectures and applications* (pp. 433-486). Hillsdale, NJ: Erlbaum.

[Zalla *et al.,* 2003] T. Zalla, P. Pradat-Diehl, and A. Sirigu. (2003). Perception of action boundaries in patients with frontal lobe damage. *Neuropsychologia, 41*, 1619-1627.

Modelling Primate Task Learning Requires Bad Machine Learning

Joanna J. Bryson
University of Bath
Artificial models of natural Intelligence
Bath, BA2 7AY United Kingdom
j.j.bryson@cs.bath.ac.uk

Jonathan C. S. Leong
Harvard University
Primate Cognitive Neuroscience
Cambridge MA 02138, USA
jleong@fas.harvard.edu

Abstract

We present a model of transitive inference which is able to account for the performance of monkeys and children on *three-item* transitive inference tasks. We do this using a modular multi-layer neural architecture which does not integrate error across layers. This system gets trapped in local minimae, and in so doing, generates errors much like those seen in monkeys and children.

1 Introduction

Transitive inference (TI) is the process of reasoning whereby one determines that if, for some quality, $A > B$ and $B > C$, then $A > C$. In some domains, such as integers or heights, this property will hold for any A, B or C. For other domains, such as sporting competitions and primate dominance hierarchies, the property does not necessarily hold (Wright, 2001). Transitive inference has become a significant benchmark task for psychologists of both animal and human cognition and has also attracted a large number of modelling attempts.

While it is well-known that the errors we make often tell us more about the nature of cognitive processes underlying behaviour than active performance does, relatively few models of transitive performance account for *failures* to learn this task. There are two sorts of failures to be accounted for. First, many subjects (both human and animal) fail to meet criteria on these tasks despite careful training. Second, though more controversially, there is a set of data due originally to McGonigle and Chalmers (1977) showing that both children and animals fail in systematic ways to generalise their ability to perform transitive 'inference' in the context of two items to a context of three items.

This paper presents a model that explains both types of errors, and proposes testable predictions on its own validity. Ironically, our work indicates that the problem with previous AI models of transitive inferences is that they learn *too well* to be appropriate models of animal behaviour. Machine learning has developed techniques that conquer a problem known as finding local minima – the problem of being attracted to a solution that, while better than the most similar and therefore obvious alternative solutions, is not actually the optimal solution. However, our research indicates that real primates may have more trouble ignoring these attractive locally-optimal solutions than many of their models do.

Our research supports the suggestions of (Buckmaster et al., 2004) and others that transitive inference relies on two separate learning processes: one to associate a stimulus with an action, and another to prioritise which of these paired associations is most salient in a context where more than one could be applied. We call our model the two-tier model, because we dedicate one tier of associative learning to each problem. The difference between our model and previous multi-layer models of transitive inference (e.g. De Lillo et al., 2001; Frank et al., 2003) is that these models use a technique called backpropogation for ensuring optimal learning across the full system. Our system keeps the two learning systems relatively independent, and so succeeds in failing where the other systems have learned too well to fully explain the data.

2 The Task

2.1 Transitive Inference and Performance

Piaget first described TI as an example of concrete operational thought (Piaget, 1954). That is, children become capable of TI when they become capable of mentally performing the physical manipulations they would otherwise use to determine the correct answer. For TI, this manipulation involves ordering the objects into a sequence using the rules $A > B$ and $B > C$, and then observing the relation between A and C.

Yet Piaget was also aware of an 'automatic' transitive performance, distinguished from true TI by the subject's ability to explain their performance (Piaget, 1928; Wright, 2001). Since the 1970s, TI has been demonstrated in young children (Bryant and Trabasso, 1971) and a variety of animals — monkeys (McGonigle and Chalmers, 1977), rats (Dusek and Eichenbaum, 1997) and even pigeons (Fersen et al., 1991) — not normally ascribed with concrete operational abilities. Siemann and Delius (1993) have shown that human adults who learned to choose between pairs of doors during an an exploration-based computer game, showed no performance difference between individuals who formed explicit transitive models and those who did not ($N = 8$ vs. 7 respectively). These results cast doubt upon the belief that all transitive *performance* (TP), choosing A over C given the knowledge that $A > B$ and $B > C$, is dependent on logical *inference*.

2.2 Characteristic TP Effects

Besides the 'inference' itself, transitive performance is characterised by a number of attributes which have generally been taken to indicate something about the processing underlying the ability (Bryant and Trabasso, 1971). Some researchers have called some of the effects into question, particularly the temporal aspect of the symbolic distance effect (McGonigle and Chalmers, 1992; Rapp et al., 1996). Nevertheless, the following effects have been shown broadly across experimental subjects, including children, monkeys, rats, pigeons, or adult video-game players (Wynne, 1998).

- *The End Anchor Effect*: subjects make an evaluation faster and more accurately when a test pair contains one of the ends. This is usually explained by the fact they have learned only one behaviour with regard to the end points (e.g. nothing is $> A$).
- *The Serial Position Effect*: even taking into account the end anchor effect, subjects do more poorly the closer the items displayed are to the middle of the sequence.
- *Symbolic Distance Effect (SDE)*: even compensating for the end anchor effect, the further apart on the series two items are, the faster the subject makes the evaluation. This effect is generally taken to contradict any step-wise chaining model of transitive inference (i.e. Piaget's concrete operations) since distant items would require *more* steps and therefore a longer reaction time (RT), not a shorter one.

2.3 Training Subjects for TP

Training a subject to perform transitive inference is not trivial. Subjects are trained on a number of ordered pairs, typically in batches. Because of the end anchor effect, there must be at least five items ($A \ldots E$) to clearly demonstrate transitivity on just one untrained pair (BD). Seven or more items would give further information, but training for transitivity is notoriously difficult. Even children who can master five items often cannot master seven. This is true even for simple sorting of conspicuously ordered items such as posts of different lengths (McGonigle and Chalmers, 1996). Normally, though, stimuli are labelled in a deliberately non-ordinal way, such as by colour or pattern, and controlled by varying the assignment of rank by subject (e.g. one subject may learn $blue < green < brown$ while another $brown < blue < green$).

The subjects are first taught to use the testing apparatus; they are presented with an object and rewarded for selecting it. Next, they are trained on the first pair DE, where only one element, D is rewarded[1]. When the subjects reach criterion, they are trained on CD. After all pairs are trained, there is generally a phase of ordered repeated training on all the pairs, but with fewer exposures per pair, which is then followed by a period of random presentations of training pairs (See phases P1–P3 in Table 1).

Once a subject has been trained to criterion, they are exposed to testing pairs. In testing, either choice is rewarded in an effort to minimise the effects of further training. The reason either stimulus should be rewarded is because whatever item is chosen is the one the subject is most likely to have expected to be rewarded for, and since learning tends to occur when expectations are violated, it is less disruptive to meet those expectations than to adopt some other reward scheme. However, the original (adjacent) training pairs are often interspersed with testing pairs during the testing phase, with the training pairs still being differentially rewarded.

[1]The psychological literature is not consistent about whether A or E is the 'higher' (rewarded) end. This paper uses A as high.

2.4 Trigram Data Sets

Table 1 finishes with a set of trigram testing. These tests are to date apparently unique to the laboratory of McGonigle, although in that lab they have been applied several times and to children as well as to monkeys. A trigram test presents three rather than two items drawn from those in the implicit sequence the subjects have been trained on. Most subjects show systematic degradation of transitive performance when exposed to trigrams.

Trigram testing has been criticised on the grounds that the sudden presence of three items might confuse the subjects and degrade their performance in itself. This criticism was addressed by McGonigle and Chalmers (1992) when they repeated their 1977 experiments to gather more data on reaction times. In 1992 they also tested their subjects on pseudo-trigrams. In pseudo-trigrams, only two classes of elements are present, one of which (chosen at random) is a duplicate (e.g. A, A, C or B, D, D). Subjects showed no performance degradation in this case. The quality of the dataset is further supported by the fact it was accounted for extremely well by the model of Harris and McGonigle (1994), which is described next.

2.5 The Production-Rule-Stack Model

Our model was originally inspired by the current best model of the trigram data set. This model is due to Harris and McGonigle (1994). They present a static, non-learning model of fully-trained subjects which accounts for the trigram data, both in aggregate and as an explanation of individual differences between subjects. The model was originally developed by Harris to account for the McGonigle and Chalmers (1977) data. This work helped motivate the the McGonigle and Chalmers (1992) study, which was in turn modelled by Harris and McGonigle.

The Harris model is based on a production-rule stack. The term *production rule* comes from artificial intelligence. It is a representation which tightly associates a particular context or *sensory precondition* with an action. A *stack* is a common representation from computer science. As the name suggests, it is a set of objects which have to be visited in order: the top item must be looked at before you can see the second-from-top item and so on. With a production-rule stack, productions are checked in order beginning from the top of the stack. If, when checked, a production's precondition is met by the environment then the second half of that rule — the action associated with the stimulus — is expressed. For example, a precondition might be *able to see* A and an action might be *grab the item holding your visual attention*.

Table 1: Phases of training and testing, taken from Chalmers and McGonigle (1984, pp. 359–360).

	Training and Criteria
P1	Each pair in order (ED, DC, CB, BA) repeated until 9 of 10 most recent trials are correct. Reject if requires over 200 trials total
P2a	4 of each pair in order. Criteria: 32 consecutive trials correct. Reject if requires over 200 trials total
P2b	2 of each pair in order. Criteria: 16 consecutive trials correct. Reject if requires over 200 trials total
P2c	1 of each pair in order. Criteria: 30 consecutive trials correct. No rejection criteria.
P3	1 of each pair randomly ordered. Criteria: 24 consecutive trials correct. Reject if requires over 200 trials total
T1	Bigram tests: 6 sets of 10 pairs in random order. Reward unless failed training pair.
T2a	As in P3 for 32 trials. Unless 90% correct, redo P3.
T2	Trigram tests: 6 sets of 10 trigrams in random order, reward for all.
T3	Extended version of T2.

The Harris and McGonigle production-rule-stack model requires the following assumptions:

1. The subject knows a set of rules of the nature "if A is present, select A" or "if D is present, avoid D".
2. The subject has a prioritisation of these rules.

For an example, consider a subject with the stack:

1. (A present) $\Rightarrow$ select A
2. (E present) $\Rightarrow$ avoid E
3. (D present) $\Rightarrow$ avoid D
4. (B present) $\Rightarrow$ select B

Here the top item (1) is assumed to have the highest priority. If the subject is presented with a pair CD it begins working down its rule stack. Rules 1 and 2 do not apply, since neither A nor E is present in the presented pair. However, rule 3 indicates the subject should avoid D, so consequently it selects C. Priority is critical. For example, for the pair DE, rules 2 and 3 give different results. However, since rule 2 has higher priority, D will be selected.

Harris and McGonigle make one more critical assumption:

3. When there are more than two items (as in the trigram test cases), an 'avoid' rule results in random selection between the items not currently attended to.

For example, consider the situation where there are three blocks available B, C, E. If the agent is applying the rule 2 above, and has found and attended to a block E, the 'avoid' action means that it is equally likely to actually grasp either B or C. This assumption explains the performance degradation of children and monkeys on the trigram data. This is why trigrams can be used to discriminate ordered-action-associative models from conventional sequential models on the basis of expressed behaviour.

Harris and McGonigle model the conglomerate monkey data so well that there is no significant difference between the model and the data. For example, over all possible trigrams, the rule-stack hypothesis predicts a distribution of 0, 25% and 75% for the lowest, middle and highest items. True inference of course predicts 0%, 0% and 100%. The squirrel monkeys in McGonigle and Chalmers (1977) showed 1%, 22% and 78%. Further, Harris was able to match the individual performance of most monkeys to a particular stack.

Without trigram data, there would be no way to discriminate which rule set the monkeys use. However, with trigram data, the stacks are distinguishable because of their errors. For example, the stack:

1′. (A present) $\Rightarrow$ select A
2′. (B present) $\Rightarrow$ select B
3′. (C present) $\Rightarrow$ select C
4′. (D present) $\Rightarrow$ select D

would always select B from the trigram BCD by using rule 2′, while the previous stack would select B 50% of the time and C 50% because it would base its decision on rule 3.

There are 8 discernible correct rule stacks of three rules each which will solve the initial training task. There are actually 16 correct stacks of four rules, but trigram experiments cannot discriminate whether the fourth rule selects or avoids (Harris and McGonigle, 1994, p.325).

3 Model

Our model is a two-tier system: one tier to learn the appropriate action associated with each stimulus, and one to learn the ordering (or *prioritisation*) of these associated 'rules' (see Figure 1). For both these learning tasks, priority is represented with a weight, the value of which is learned through reinforcement. Larger weights correspond to higher priorities. Which 'rule' is fired is determined first by which stimulus present is associated with the largest weight, and then by which actions in that stimulus' associated action list has the largest weight.

The two-tier model contains of a list of perceptual categories corresponding to the different stimuli seen. Half of the agent's task is to prioritise this list. The other half of the task (the second tier of the model) is prioritising the associations

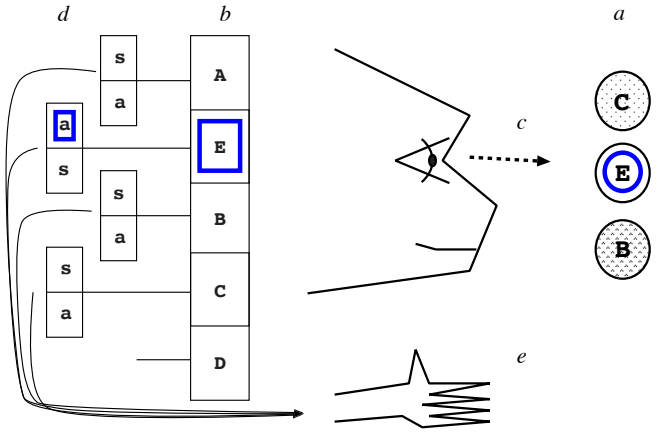


Figure 1: The two-tier model. When the agent observes a set of stimuli (a), a weight vector (b, the first tier) determines which item present is most salient. This attracts visual attention (c) and determines which rule vector (d, the second tier) selects the appropriate action (**s**elect or **a**void). This determines what item the agent grasps (e). The two vectors that were most recently active (b and one of d) are then updated in response to the reward as per Equation 1.

associated with each stimulus. The same learning rule is applied to both (see below.)

Figure 1 illustrates how a subject chooses a stimulus under the two-tier model. At the beginning of a trial, the subject is presented with some number of stimuli, generally two or three. If any of the stimuli are novel, they are added to the first tier by a process described below. Next, the subject focuses visual attention on the stimuli present in the scene with the highest priority. The highest-priority stimulus is the one associated with the largest weight. Next, the subject applies the highest priority action in the action array from the second tier which is associated with that stimulus. The subject either selects the object it is attending to, or 'avoids' that object by grasping another object. If there is more than one other object present and the subject is avoiding, then the unattended object chosen is determined at random. For either tier, in the case where more than one eligible tier element has the same priority, one of these options is selected at random.

After a stimulus is selected, the subject is either rewarded or not. Weights for both tiers are updated independently after every trial. We use a simple step function which roughly approximates known conditioning models (Waelti et al., 2001; Rescorla and Wagner, 1972).

The same learning rule is used for both tiers. All of the weights in a single list are normalised, that is they always sum to 1. When a new stimulus is seen, it receives the weight $1/N$, where N is the current number of distinct stimuli categories so far seen. New items in the stimulus list are further associated with a new action list, which is initialised with two actions, *select* and *avoid*, both of which are given a starting weight of 0.5.

The weights for any particular list (the stimuli list in the first tier or one of the action lists associated with a single stimulus in the second tier) are represented as a vector $\mathbf{w}$. Consider the pair XY, where X is the list element the subject attended to and Y is a near alternative[2], then $\mathbf{w}_X$ and $\mathbf{w}_Y$ are the weights associated with X and Y respectively. The update rule for these weights is this:

$$\text{If X is correct and } \mathbf{w}_X - \mathbf{w}_Y < \tau\text{), add } \delta \text{ to } \mathbf{w}_X\text{;}$$
$$\text{else, if X is incorrect, add } \delta \text{ to } \mathbf{w}_Y\text{.} \qquad (1)$$

where τ and δ are free parameters, held constant for any particular subject, but varied across subjects for the experiments. τ is a threshold, over which reward is so expected that it no longer prompts learning (Waelti et al., 2001). δ is the amount a weight is changed by a single bout of learning. If weight change occurs, $\mathbf{w}$ is subsequently renormalised.

The goal of this model is *not* to try to to improve on the Harris and McGonigle (1994) outcome for modelling the trained monkeys, since that is already excellent. Rather, the goal is to build a model which *learns* such a model in a biologically plausible way, and to thus better understand what sort of representation might be underlying the transitive performance shown in squirrel monkeys and young children. Thus our model succeeds if it can learn from the same experience the primate subjects are exposed to a model which behaves exactly as the production-rule-stack model does.

4 Methods

4.1 Modelling through Artificial Life

The experiments in this paper were performed in artificial life (ALife) simulations. ALife is often a very intuitive way to express a cognitive hypothesis. In addition, it can allow researchers to evaluate algorithms that are too complex to analyse formally (Axtell, 2000). ALife evaluations operate by running simulations, then performing standard hypothesis testing to see whether the simulated results are a good match to the original data.

An ALife model can be thought of as a very well specified hypothesis. Once a model has been built, the process of simulation also allows one to search broad parameter spaces that would be relatively difficult or expensive to test in the laboratory. These runs then serve as predictions from the model. If desired, a relatively sparse set of these predictions, perhaps those that are most surprising or vary most between different versions of the hypothesis, can then be tested against experiments with living subjects.

[2] If X is a stimulus, Y is one of the other stimuli present chosen randomly, unless the rule was *avoid* in which case it is the item actually grasped. If X is an action then Y is the second-highest priority action for that stimulus.

Further validation of an ALife model occurs when simulation results unexpectedly converge with previously-observed, real-world phenomena. This was the case for our experiments.

5 Results and Discussion

5.1 Experiment 1

Procedure

The first experiment was essentially a pilot experiment. In this condition we did not use the full two-tier architecture, but rather tested the learning algorithm shown in Equation 1 on a single-tier model. For this experiment, there was only a single vector with one element for each stimulus. The agent would choose the stimulus corresponding to the vector element with the highest priority weight.

We also did not use the full training regime shown in Table 1, but rather simply exposed the subjects to all training pairs in a random order for approximately 400 trials. This training regime is sometimes used with rats (Wynne, 1998).

Results and Discussion

The single-tier system learns to order the stimuli perfectly 100% of the time, provided that $1/N \geq \tau$. The last error by these systems is normally made by the 100^{th} trial, while weights stop changing (or *stabilise*[3]) about 50 trials later.

This perfect ability to pass criteria is very unlike primates, which normally need a training regime. Further, this model performs perfectly on trigrams, again unlike real primates, since once the ordering is leaned it will always select the highest priority stimulus.

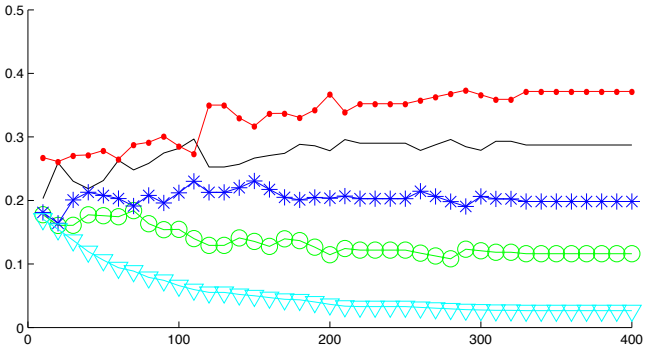


Figure 2: A typical result for a one-tier learning agent. X-axis: trial number; Y-axis: weights of the stimuli vector (sum to one). Free variables are set to Parameters: $\tau = .08$, $\delta = .02$. Key: $A\bullet$, $B-$, $C*$,$D\bigcirc$, $E\bigtriangledown$.

Figure 2 shows a single exemplar of a typical result where $1/N \geq \tau$. If $\tau > 0.1$, then a stable solution for five items cannot be reached. This is because there is no way that five weights can be more than 0.1 different from each other, yet sum up to 1 (see Equation 1 and the discussion of normalisation.) If $\tau = 0.1$, then it is possible that the weights can be $[0, 0.1, 0.2, 0.3, 0.4]$ which does sum to 1. For any value of $\tau < 0.1$ there are many possible stable solutions.

[3]Normally in AI, agents are considered to have fully learned a task only when their weights have stabilised. This of course can't map directly to the animal research, where learning must be judged by expressed behaviour.

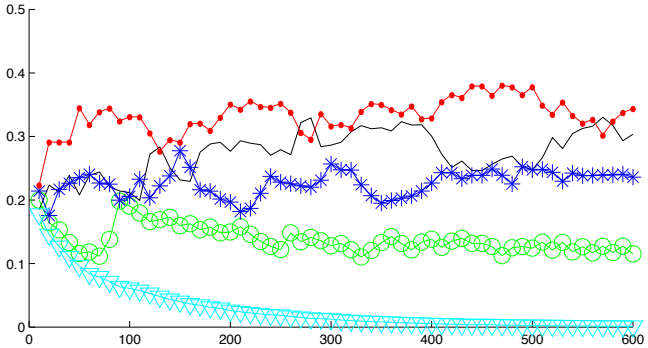


Figure 3: One-tier learning in an agent with a 'stupider' parameter set: $\tau = .12$, $\delta = .02$. This cannot find a stable solution (see text) thus occasionally gives wrong responses. Key: $A\bullet$, $B-$, $C*$,$D\bigcirc$, $E\bigtriangledown$.

If learning can't stabilise, the model is open to a 'hot hands'-like phenomenon (Gilovich et al., 1985), where a solution that has recently been very successful may get more favour than it deserves. When there is a chance reiteration of one particular pair, the higher element of that pair can accumulate so much reinforcement that its weight surpasses the element that should be above it. Thus the agent over-estimates the value of an item because of its recent "winning streak". This is illustrated in Figure 3. However, notice that even so, these agents learn the task very quickly and make very rare mistakes, so would easily pass behavioural criteria.

These results do provide one possible explanation for individual differences in transitive task performance. Individual differences in stable discriminations between priorities can affect the number of items that can be reliably ordered.

5.2 Experiment 2

Procedure

In the second condition we used the full two-tier model, but still trained it simply by presenting training pairs in random order. We tested learning in the two-tier model across a range of parameter values: every combination of τ drawn from .08, .1, .12, .14 and δ drawn from .01, .02, .04, .08, .12, .16. 12 subjects were run with each possible parameter combination for a total of 288 subjects.

Results and Discussion

Without a training regime, only a fifth of two-tier agents learn the training pairs entirely successfully (56 of 288, see Table 2, column 2 below.) Those agents that do learn the training pairs successfully perform on trigram testing exactly as described by (Harris, 1988), because a snapshot instance (that is, one with learning frozen) of a successfully trained two-tier model is logically equivalent to a production-rule stack.

Figure 4 illustrates a typical (failing) exemplar outcomes for this case. Rule selection is made very early in training and remains stable once established, so is not represented

in these figures except in the captions. Nevertheless, the added complication of rule learning defeats the simple training regime used in the one-tier experiments. With this training regime, the two-tier model generally learns either the solution shown in Figure 4 or a symmetric one with the select (B) and avoid(C) fighting for top priority.

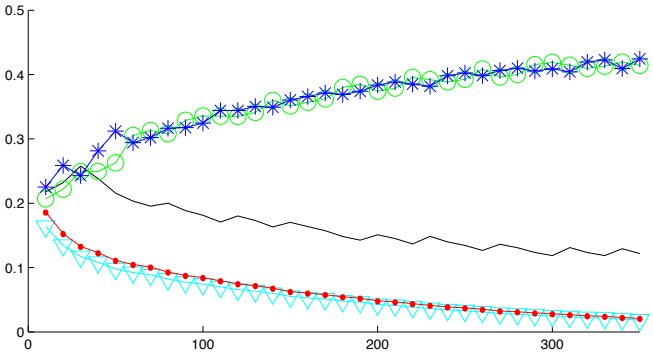


Figure 4: Two-tier learning with no training regime. Rules learned in descending order of priority: select $D\bigcirc$, avoid $C*$, avoid $B-$. The system cannot stabilise because it is far from a complete solution, but it behaves correctly for every training pair except CD. Parameters: $\tau = .08, \delta = .02$.

There are only a limited number of accurate solutions the two-tier model can have, corresponding the correct rule stacks enumerated by Harris (1988). A correct solution must be either an ordered sequence of selects $[s(A)s(B)s(C)]$, a reverse order sequence of avoids $[a(E)a(D)a(C)]$ or an ordered cross of these (e.g. $[s(A)a(E)s(B)]$). See Table 2, column 1 below for a complete list.

Although the rules learned by the typical (failing) two-tier agents without training regimes have a very different order, they still perform well on the training task. For each failing agent, only one training pair is incorrect: that containing the two top-priority stimuli. For example, in Figure 4, the only training pair which cannot be handled is BC. Notice that although the weights in Figure 4 has not stabilised, the behaviour has. Whether $a(B)$ or $s(C)$ is highest priority, when presented with B, C, the agent will (incorrectly) grasp C. These agents display something like the Serial Position Effect by confusing only central pairs. Taken in aggregate, the agents display the full SPE: the most frequent errors involve the two central pairs, but there are occasionally errors involving other elements.

All the agents show the End Anchor Effect. Agents quickly learn rules which avoid making errors involving the two end points. The agent in Figure 4 may appear to neglect the two endpoints, since the weights of the stimulus-rule pairs associated with those stimuli are very low. But in fact, this agent gets the end pairs (AB and DE) correct 100% of the time by associating the rule *avoid* with B and *select* with D. By reducing the values of the A and E rules, the agent that learned the 'correct' behaviour in the end-point cases. Associating this knowledge with the inner member of the end pair protects the agent from a possible incorrect rule associated with the outer member. However, this association leaves such agents with no possible means to correctly learn both middle pairs. The learning system winds up fixated on trying to solve an impasse in the middle of the sequence, but the learning algorithm, based on gradual change, cannot solve that quandary.

19% of the time two-tier agents without a training regime *do* learn a correct solution. If successful learning was the simple consequence of the agents being at chance for learning a rule about either the inner or the outer element of the two end pairs, we would expect that the agents would learn both ends correctly 25% of the time, each end 25% of the time, and neither end 25% of the time. We can dismiss this as the full explanation for the agent's failure: $\chi^2(3, N = 288) = 35.68, p < .001$. The fact that the inner end-pair elements, B and D occur in twice as many pairs as the end element leads the two-tier model to both inner cases 40% of the time (166 in 288), not 25%. When correct solutions were learned, they came evenly from all parameter values, and seemed evenly distributed across all possible correct solutions (Table 2, column 2).

We would obviously like to compare these results with the outcomes of primate subjects who fail to meet criterion on the initial training for transitive learning. Although no trigram results were reported for monkeys or children that missed criterion, one monkey subject, Roger, passed criterion but still showed a consistent error between the 3^{rd} and 4^{th} item (Harris and McGonigle, 1994, p. 332). Roger's errors are in keeping with the results of this model.

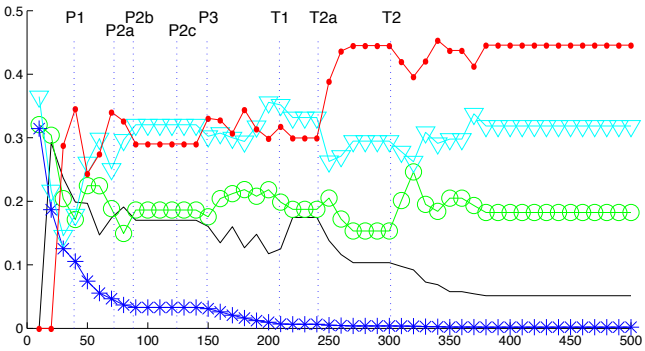


Figure 5: Rule Learning with Phased Training. Labelled lines indicate the *end* of training and testing phases (see Table 1). This agent arrives at different stable solutions at different points, but they are all correct. Here the rules are: select $A\bullet$, avoid $E\bigtriangledown$, avoid $D\bigcirc$. The agent succeeds with very 'stupid' parameters: $\tau = .12, \delta = .06$.

5.3 Experiment 3

Procedure

In the third condition, we trained two-tier models using the regime in Table 1. We ran these test over the same range of parameters as for the previous condition, again with 12 instances of each, for a total of 288 subjects.

Results and Discussion

Results, shown in Table 2, are that 88% (254 of 288) of agents successfully learn the training pairs and therefore the TP task.

Table 2: Production-rule-stack equivalents to solutions by monkey subjects and by two-tier agent subjects undergoing various forms of training. The distribution of solutions is strongly determined by the order training pairs are presented. The analysis of the monkeys' correlated stacks was performed by (Harris and McGonigle, 1994).

Correct Stacks	No Regime	Regime starting w/ E, D after training	Regime starting w/ E, D after testing	Regime starting w/ A, B after training	Regime starting w/ A, B after testing	Starting w/ A, B McGonigle & Chalmers
s(A) s(B) s(C)	8	51	41	–	–	–
s(A) s(B) a(E)	12	68	26	–	–	–
s(A) a(E) a(D)	3	–	1	4	2	2
s(A) a(E) s(B)	7	4	16	3	1	2
a(E) a(D) s(A)	9	–	1	57	50	–
a(E) a(D) a(C)	8	–	–	59	47	1
a(E) s(A) a(D)	7	3	–	4	11	–
a(E) s(A) s(B)	2	1	13	–	3	–
Total Correct	56	127	98	127	114	5
Total	288	144	144	144	144	7

The two-tier model is somewhat more like the squirrel monkeys than the children in that it more reliably learns TP than children given the same training regime. Nearly all successful agents converge quickly, and the ones that fail to meet criteria fail early, usually by Phase 2a.

Successful learning for agents with phased training is highly dependent on δ; when δ was large (values in .08, .12, .16), one in four agents failed, otherwise there were only a very few failures (3), all of which had the lowest tested δ, $\delta = .01$ ($N_{\delta=.01} = 48$). Since δ determines the rate of change after training, it is unsurprising that a very low δ results in a slow learner. Even when such agents make criterion, they may never learn a stable solution (see e.g. Figure 6). Interestingly, agents do sometimes learn when $\delta > \tau$, which means that for any trial on which learning occurs, the attended item will change places in the priority stack.

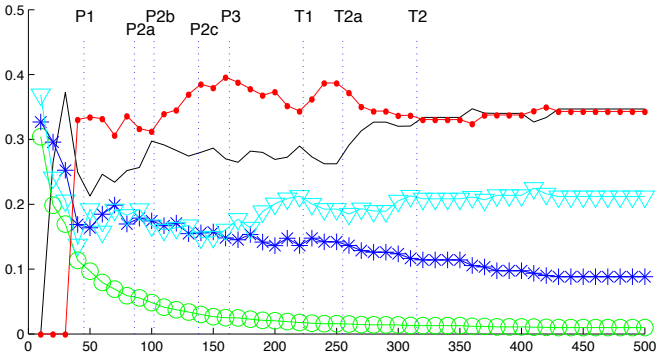


Figure 6: Phased Training where learning slips during trigram testing. Rules are: select $A\bullet$, select $B-$, and either avoid $E\triangledown$ or select $C*$. Parameters: $\tau = .12$, $\delta = .02$

One advantage of the two-tier model over a single learning tier can be seen in fact that, during initial training, stable representations are learned that involve rules with nearly the same priority. This is because some rules will never be compared. For example, items E and A do not occur together in any of the training pairs, so there is no pressure to differentiate their weights prior to TI testing (see Figure 5, phases P2b–P3). This is significant if neural systems have limited capacities to discriminate different stable orderings (see discussion for Experiment 2, and Cowan, 2001; Bryson and Lowe, 1997; Gallistel et al., 1991). For tasks involving stimuli which never co-occur, the rule representation allows for stable learning with either more items or larger values of τ. A more natural example of such a task than TI would be navigation, where some landmark features might never occur in the same place.

Another thing to notice in the phased learning results is that significant learning occurs during testing. This phenomena was also reported with monkeys (Harris and McGonigle, 1994). Learning occurs because rules that previously were never compared (e.g. those triggered by any two non-adjacent items) will be now. If their weights do not already happen to be at least τ apart, learning is triggered, regardless of whether they were correct or how they are reinforced. This explains the utility of continuing to differentially reinforce training pairs, a common procedure during the testing phase of the TI task.

6 Conclusions and Predictions

The results of our model have not only met but exceeded the goals of our simulation. We have achieved our goal by showing that a system like that of Harris and McGonigle can be learned, and with learned with a simple, biologically-plausible learning algorithm. The model exceeds our original goals by displaying the End Anchor and Serial Position effects, and by requiring the same phased training that children and squirrel monkeys require to have a similar number of agents pass criteria. That these features of the model were unintended consequences of the two-tier structure further validates both our model and the work of Harris and McGonigle (1994).

Our model makes a number of testable predictions:

- Visual attention should settle on the item associated with a rule just before the grasp is made. In the case of an *avoid*, this would not be the item that is selected.

- In general, reaction times and visual scanning behaviour should be different for select and avoid rules.
- If individuals who fail to pass criteria on training pairs are given trigram testing, most should show a misordered priority stack with high priority rules for neighbouring pairs of non-endpoint stimuli.
- For individuals, the ordering of a newly presented item (as in de Boysson-Bardies and O'Regan experiment 3) should be determined by the existing rule stack. For example, if the rule stack is all selects as in Eq. 2.5, a new item would be positioned last or second to last, if they were all avoids it would be positioned first.

Testing these predictions requires running trigram experiments after TP pair training in order to discriminate which rules were learned by individual subjects. Our current work includes a collaboration intended to test most of these predictions. We are also working on furthering the biological plausibility of the two-tier model, including extending it to account for the Symbolic Distance Effect.

References

Axtell, R. (2000). Why agents? On the varied motivations for agent computing in the social sciences. Technical Report 17, Brookings Institute: Center on Social and Economic Dynamics, Washington, D.C.

Bryant, P. E. and Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature*, 232:456–458.

Bryson, J. J. and Lowe, W. (1997). Cognition without representational redescription. *Behavioral and Brain Sciences*, 20(4):743–744. Commentary on Ballard et al. "Deictic codes for the embodiment of cognition".

Buckmaster, C. A., Eichenbaum, H., Amaral, D. G., Suzuki, W. A., and Rapp, P. R. (2004). Entorhinal cortex lesions disrupt the relational organization of memory in monkeys. *The Journal of Neuroscience*, 24(44):9811–9825.

Chalmers, M. and McGonigle, B. O. (1984). Are children any more logical than monkeys on the five term series problem? *Journal of Experimental Child Psychology*, 37:355–377.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Brain and Behavioral Sciences*, 24(1):87–114.

de Boysson-Bardies, B. and O'Regan, K. (1973). What children do in spite of adults' hypotheses. *Nature*, 246:531–534.

De Lillo, C., Floreano, D., and Antinucci, F. (2001). Transitive choices by a simple, fully connected, backpropagation neural network: implications for the comparative study of transitive inference. *Animal Cognition*, 4(1):61–68.

Dusek, J. A. and Eichenbaum, H. (1997). The hippocampus and memory for orderly stimuls relations. *Proceedings of the National Academy of Sciences, USA*, 94:7109–7114.

Fersen, L., Wynne, C. D. L., Delius, J., and Staddon, J. E. R. (1991). Transitive inference formation in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 17:334–341.

Frank, M. J., Rudy, J. W., and OReilly, R. C. (2003). Transitivity, flexibility, conjunctive representations, and the hippocampus. ii. a computational analysis. *Hippocampus*, 13(3):341–354.

Gallistel, C., Brown, A. L., Carey, S., Gelman, R., and Keil, F. C. (1991). Lessons from animal learning for the study of cognitive development. In Carey, S. and Gelman, R., editors, *The Epigenesis of Mind*, pages 3–36. Lawrence Erlbaum, Hillsdale, NJ.

Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive Psychology*, 17:295–314.

Harris, M. R. (1988). *Computational Modelling of Transitive Inference: A Micro Analysis of a Simple Form of Reasoning*. PhD thesis, University of Edinburgh.

Harris, M. R. and McGonigle, B. O. (1994). A model of transitive choice. *The Quarterly Journal of Experimental Psychology*, 47B(3):319–348.

McGonigle, B. O. and Chalmers, M. (1977). Are monkeys logical? *Nature*, 267:694–696.

McGonigle, B. O. and Chalmers, M. (1992). Monkeys are rational! *The Quarterly Journal of Experimental Psychology*, 45B(3):189–228.

McGonigle, B. O. and Chalmers, M. (1996). The ontology of order. In Smith, L., editor, *Critical Readings on Piaget*, chapter 14. Routledge, London.

Piaget, J. (1928). *Judgment and reasoning in the child*. Routledge and Kegan Paul, London.

Piaget, J. (1954). *The Construction of Reality in the Child*. Basic Books, New York.

Rapp, P. R., Kansky, M. T., and Eichenbaum, H. (1996). Learning and memory for hierarchical relationships in the monkey: Effects of aging. *Behavioral Neuroscience*, 110(5):887–897.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. H. and Prokasy, W. F., editors, *Classical Conditioning II*, chapter 3, pages 64–99. Appleton, New York.

Siemann, M. and Delius, J. D. (1993). Implicit deductive reasoning in humans. *Naturwissenshaften*, 80:364–366.

Waelti, P., Dickinson, A., and Schultz, W. (2001). Dopamine reponses comply with basic assumptions of formal learning theory. *Nature*, 412:43–48.

Wright, B. C. (2001). Reconceptualizing the transitive inference ability: A framework for existing and future research. *Developmental Review*, 21(4):375–422.

Wynne, C. D. L. (1998). A minimal model of transitive inference. In Wynne, C. D. L. and Staddon, J. E. R., editors, *Models of Action*, pages 269–307. Lawrence Erlbaum Associates, Mahwah, NJ.

Modelling Perceptual Phenomena using Temporal Abstraction Networks

Neil Madden and Brian Logan

University of Nottingham
School of Computer Science and Information Technology
Nottingham, UK
{nem,bsl}@cs.nott.ac.uk

Abstract

We present Temporal Abstraction Networks, a novel cognitive architecture which can be used to model a variety of perceptual phenomena. The architecture is based on processes operating on collections of time-limited buffers in a parallel model of cognition and draws on aspects of the Multiple Drafts theory [Dennett and Kinsbourne, 1992]. We briefly describe the architecture and show how it can be used to model two relevant experiments from the literature: colour phi [Kolers and von Grünau, 1976], and the cutaneous "rabbit" [Geldard and Sherrick, 1972].

1 Introduction

Modelling cognitive phenomena in which the time of perception plays a role is an important challenge for cognitive science. A perceptual event is by its very nature transient. In order to reason about perceptual events we either have to extract information from them as they occur, or try to recreate details from causal evidence after the fact. A number of experimental studies, e.g., [Kolers and von Grünau, 1976; Geldard and Sherrick, 1972] have suggested that interpretations of events can override direct sensory evidence. For some sequences of perceptual events of short duration, the interpretation of individual events in the sequence depends on the characteristics of the sequence as a whole. This 'backwards referral in time', in which later events influence the perception of earlier events, is difficult to account for within a serial model of cognition without incorporating implausible delays (basically delaying sensory experience "until all the data are in").

Dennett and Kinsbourne [1991; 1992] have proposed the *Multiple Drafts* theory as a way of modelling such cognitive processes. The Multiple Drafts theory is based on a parallel, distributed view of cognition, in which large numbers of processes work independently on multiple interpretations of data simultaneously. These are the multiple *drafts*. Eventually a single draft may become dominant, but no draft is ever entirely safe from revision.

In this paper we present Temporal Abstraction Networks, a cognitive architecture for perceptual processing which draws on aspects of the Multiple Drafts theory. A Temporal Abstraction Network (TAN) consists of a network of inference processes working in parallel on collections of data over different temporal intervals (see Figure 2 for an example). TANs can be used to model a variety of perceptual phenomena which present difficulties for more conventional serial models of cognition, such as ACT-R [Anderson and Lebiere, 1998] or Soar [Newell, 1990; 1992]. As an illustration, we show how TANs can be used to model two perceptual phenomena that have been claimed to cause problems for serial models of cognition [Dennett, 1991]; namely, colour phi [Kolers and von Grünau, 1976] and the cutaneous 'rabbit' [Geldard and Sherrick, 1972; Geldard, 1977]. The architecture and models have been implemented using the SIM_AGENT toolkit [Sloman and Poli, 1996].

The remainder of this paper is organised as follows. In the next section we give a brief overview of the Multiple Drafts theory, focusing on its implications for modelling perceptual phenomena in reactive agents. In section 3 we introduce the time-limited buffers and buses which form the key components of the Temporal Abstraction Network architecture and describe how these can be combined to give models of reactive perceptual processing in simple agents. In section 4 we present models of Kolers and von Grünau's 'colour phi', and Geldard and Sherrick's 'cutaneous rabbit' experiments. In section 5 we briefly discuss related work before considering the implications of our approach for the Multiple Drafts theory in section 6.

2 The Multiple Drafts Theory

The Multiple Drafts theory of consciousness proposed by Dennett and Kinsbourne [1991; 1992] is an attempt to explain general cognitive processes, and how they can give rise to consciousness, without appealing to a *Cartesian theatre* — a central process where everything "comes together". Instead, the Multiple Drafts theory posits a highly parallel view of cognition where processing and interpretation are carried out in a distributed manner. Different interpretations constitute the multiple *drafts* which compete for (temporary) dominance. A draft which survives long enough can become relatively uncontested, and so become the dominant interpretation of events. However, no draft is ever entirely safe from further revision or reinterpretation.

The advantages of a parallel architecture over more traditional serial theories of cognition are most apparent when time is taken into account. An agent situated in an environment must act in a timely fashion in order to respond to events occurring in the world. However, as events continue to be perceived during the selection and performance of an action, there is often a need to revise the interpretation(s) of events to take account of new information. Rather than commit to a single interpretation of an event, and then later possibly have to backtrack or revise the interpretation in the face of new evidence, it is quicker to keep track of multiple possible interpretations of an event (multiple drafts) in parallel and simply drop those which are no longer supported by evidence. It is also preferable to be able to make decisions without delaying processing until all possible data has arrived, but without having to commit to a single interpretation of events too early. By allowing multiple drafts to exist simultaneously we can select actions based on the current dominant interpretation, while still allowing future information to revise or create new drafts which may then influence future action selection. A serial model of cognition commits us to either delaying processing or facing possible costly back-tracking (the two alternatives are characterised by Dennett as "Stalinesque" and "Orwellian" revisionism, respectively. [Dennett, 1991] p.116–24).

The cognitive architecture presented in this paper draws on several aspects of the Multiple Drafts theory. The architecture allows the formation of networks of parallel processes, with no single central process in ultimate control. Different processes work on (potentially) conflicting interpretations of events, and these drafts may persist for different lengths of time, depending on whether they are considered useful by other processes. No direct revision of drafts occurs; instead new interpretations are generated which outlast the obsolete drafts.

In this paper we focus on reactive models of perception, and do not attempt to model higher deliberative processes, or to demonstrate how reports of perceptions are assembled. In addition, we do not model long-term memory or persistence of drafts beyond the short time-scales of immediate perception.

3 The Temporal Abstraction Network Architecture

The Temporal Abstraction Network architecture consists of a set of processes that make inferences based on *symbolic* data within a certain temporal window. The processes are connected via a bus architecture, allowing the conclusions drawn by one process to form the inputs to other processes (including themselves), see Figure 1.

3.1 Time-limited Buffers

Each inference process has an *input buffer* with specified capacity and duration. The *capacity* of a buffer is the number of items that may be present in the buffer at any given time. The buffer's *duration* is the maximum length of time elements can remain in the buffer before they are forgotten. The duration and capacity of a buffer are independent of each other, e.g., a buffer may have large capacity but short duration or small capacity but longer duration. New inputs are added to the buffer in first in first out fashion — items arriving at a full buffer cause the oldest items in the buffer to be overwritten.

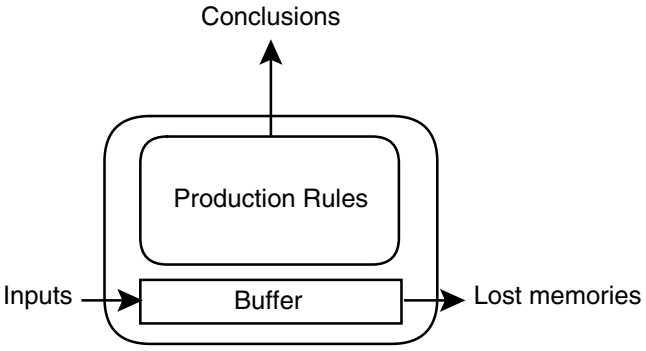


Figure 1: Conceptual model of an inference process.

Each process can draw conclusions based on the current contents of its input buffer, using an inference procedure as detailed below. Each inference process may have a different buffer duration, allowing different process to draw conclusions based on events occurring over different lengths of time. At any given point, the contents of the buffers constitute the entire state of the perceptual system.

3.2 Inference

The system as a whole runs in cycles. Each inference process contains an *inference engine*, a set of production rules that are used to spot patterns in input data, and draw conclusions based on these patterns. At each cycle each inference process's production rules are matched against the contents of the input buffer and a single rule fired. The inference engines all operate at the same rate, and the production rules have a chance to fire at each cycle regardless of whether any new inputs have arrived at the corresponding buffer. Each rule can generate a single output which is automatically transferred to the output bus of the process (see below).

More precisely, at each cycle each process:

1. Removes expired items from the buffer. An expired item is one which has been present on the buffer for longer than the buffer duration.
2. Adds any new items that have arrived since the last cycle, over-writing the 'oldest' items if the total number of items exceeds the capacity of the buffer.
3. Matches rules against current contents of the buffer.
4. Selects a single rule and runs it (if possible).
5. Transfers the output of the rule to the output bus.

Inference processes can be categorised into three main types based on the kinds of inferences they perform: processes that filter information; those that apply transformations; and those that integrate features of many items in a buffer and produce conclusions based on the properties of the composite grouping. An individual inference process may perform some combination of the above three categories of abstraction.

3.3 Buses and Information Propagation

The production rules within each process are used solely to draw conclusions based on the current state of the input buffer — they cannot alter the buffer in any way. Instead, output from the rules are passed on to a *bus*, that transmits them to the input buffers of other processes. In this way, data can be abstracted as it progresses through the network of connected processes, with different abstractions persisting for different lengths of time. We envisage that processes further up a chain (further from the initial percepts) would have buffers spanning a longer duration of time than the lower level processes, allowing the system as a whole to remember more abstract conclusions, while most of the details are forgotten.

To use the terminology of the Multiple Draft theory, the output of a process would represent a *draft* interpretation of events, and there may be multiple simultaneous drafts existing in multiple buffers at any given time. In addition, the bus connection architecture means that a single conclusion from a low-level process can be delivered to multiple higher-level processes, allowing for multiple drafts to be formed based on the same data, potentially producing different conclusions or interpretations. The many-to-many connections of the buses mean that the architecture is not limited to a rigid hierarchy with all conclusions eventually converging on a central process. There may be *local bottlenecks* where some combination of outputs are brought together for integration, but this does not imply that further integration must occur: the flow of information may diverge as well as converge.

The survival of a particular draft depends solely on its relevance to higher-level processes; drafts that are irrelevant to higher-level processes (in other words, do not match any patterns in the rules) will simply expire from the buffers or be overwritten by more recent drafts.

The architecture assumes that buses propagate information instantaneously—that no time is spent transferring information from one process to another, even over several buses. In practice, this means that an output from one process on a particular cycle will be present on the buffers of connected processes by the start of the next cycle. In the current implementation conclusions are transferred immediately (i.e., during the current cycle) but only made visible to rules at the start of the next cycle—this means order of execution of the processes is not important.

3.4 Feedback Loops and Alarms

While each process cannot write to its own buffer directly, it can do so in a round-about way, by making use of a feedback loop. Each process is connected to two buses; one for input, another for output. However, any bus can be connected to an arbitrary number of other processes, and other buses. This allows the output bus of a process to be connected to the input bus, creating a feedback loop. Feedback loops are useful for a variety of reasons, the most important of which is keeping track of information over a longer period of time than the input buffer allows. By using a feedback loop, a process can periodically (e.g. every cycle) send itself a message keeping track of important information, for instance keeping a running total of events that have occurred.

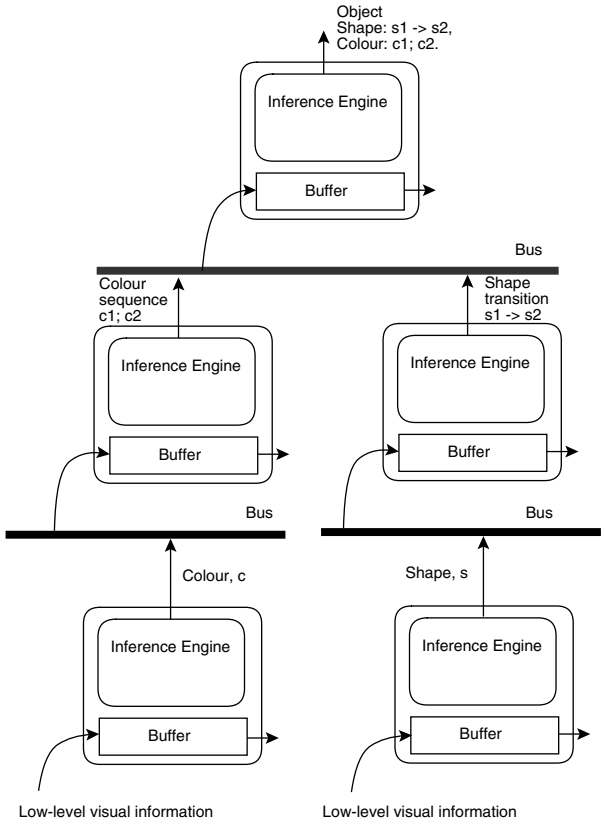


Figure 2: Network for the colour phi experiment

In addition to feedback loops, buses can be connected together in other ways, in order to provide different functions. One such function is to provide an alarm mechanism [Norman and Long, 1996; Sloman, 1998], by which particularly important events detected at a low level could bypass intermediate levels of processing and be passed directly to higher level processes for immediate action. This can be accomplished by connecting the low level bus to the high level bus via a process that acts as a filter, to determine which events require immediate attention (for instance, acting as an attention filter [Logan, 2000]).

4 Models

We have developed Temporal Abstraction Network models of a number of perceptual phenomena. In this section we briefly describe two such models: the colour phi phenomenon, and the cutaneous "rabbit", and show how the TAN architecture was used to model them. The models were implemented using the SIM_AGENT toolkit [Sloman and Poli, 1996].

4.1 Colour Phi Phenomenon

Kolers and von Grünau's [1976] colour phi phenomenon demonstrates an interesting aspect of how visual stimuli are perceived over time. In the experiments subjects were briefly shown a coloured shape at a certain position. The shape then disappeared and was swiftly replaced by another shape of a

different colour in a different position. The stimuli were typically presented for 150ms with a 50ms gap between them. A number of variations on the experiment were performed, with some keeping the colour constant and changing the shape, some changing the colour but not the shape, and others changing both. Subjects were asked to indicate how the colour and shape changed. The results show that subjects perceive shape as changing continuously between the two points, whereas colour is perceived as changing abruptly somewhere between the two points. Moreover, the colour "filled-in" the intermediate shapes.

The model that we have created discriminates colour (a property of surfaces) separately from shape (a property of edges) and also discriminates changes in colour separately from changes in shape. The colour processes interpret a change in colour as an abrupt change (or rather, make no interpretation at all of changes in colour), while shape processes perceive changes as continuous (a single object changing shape). We believe this is a reasonable dichotomy, as objects in the natural world often vary in apparent shape over time (e.g. as a result of relative motion), but rarely change in hue. This is also consistent with findings that colour plays only a small role in perception of movement and other properties of objects [SFN, 2005]. Further processes then try to integrate this change data into a conclusion of the form "shape s_1 changing to s_2 AND colour c_1 followed by c_2". No "filling-in" of intermediate stages is performed at all by the perceptual processes. Again, this is reasonable, as we have reached a level of abstraction at which the agent reasons about *behaviour* of objects over time, rather than the properties of individual perceptions of the objects. This is a more useful level of abstraction for making predictions about the future behaviour of objects. By identifying objects and making predictions about their movements, an agent can anticipate and plan ahead rather than simply responding to events as they occur. The use of short-duration buffers allows the agent to aggregate information from several events in succession and spot overall trends in data, while the bus architecture means that the use of these buffers does not preclude the propagation of such information to other processes for immediate action. Once an overall trend is spotted, it is not necessary to recall the original data for revision, or to spend time "filling-in" the missing pieces. Instead, it is sufficient to generate a conclusion describing the overall trend. It is only when explicitly asked to describe the events (when asked to form a report) that an attempt is triggered to recreate the (phantom) intermediate stages.

Figure 2 shows the network of inference processes used to model this experiment. The two lowest-level processes extract shape and colour information from low-level visual sensory data. The details of how these processes operate is dependent on the underlying representations of visual data, and are omitted. The next layer of processes are also split into a colour/shape divide. Each process in this layer has a longer duration buffer than the lower level processes, and aggregates information about colours/shapes over that time period. The buffers can vary in duration independently of each other, but the results from colour phi suggest that a duration of at least 50ms is required for both. The outputs of these intermediate levels express a *change* in the underlying property. This change is encoded as a start and end state together with an indication of the sort of change (continuous transition or abrupt change). The final process at the top of the figure integrates shape and colour information to describe the recent changes in properties of an object. This information can then be used by further processes to make predictions and decide upon appropriate actions. The intermediate processing and raw sensory data are not typically made available to other processes (although they could be), and so this information disappears when it expires from the short-duration buffers in the low-level processes. Thus, when a higher-level process wishes to generate a report of the episode, it only has the high-level interpretation of a continuously changing shape and an abrupt colour change.

4.2 Cutaneous "Rabbit"

Geldard and Sherrick's cutaneous "rabbit" experiments [1972; 1977] offer evidence for another perceptual phenomenon called *sensory saltation*. In the experiments a series of short 'taps' (of about 2ms duration) were delivered to the arm of a subject. The taps are delivered with intervals of between 0–500ms. In the original experiment the taps were delivered in sequences to different locations on the arm — for instance, five taps at the wrist, followed by five between wrist and elbow, and then five more at the elbow. Subjects reported that the taps had been more or less evenly spaced along the arm — as if a little rabbit was hopping up the arm. This effect is illustrated in Figure 3. Variation in the interval between taps (inter-stimulus interval, ISI) causes differences in the effect felt. If the ISI exceeds approximately 200 ms then the effect is not felt; the taps are felt at their correct locations. With an ISI of 20 ms or less the number of taps felt becomes illusory, for example, 15 taps may be perceived as just 6. Inter-stimulus intervals between these extremes cause variations in the apparent spacing and intensity of the displaced taps, but an overall sensation that the taps were more or less evenly spaced between the location of the first tap and that of the last.

A model of this experiment is shown in Figure 4. The model uses a feedback loop to aggregate information about individual taps into information about a sequence of taps. As in the colour phi model, this aggregation allows the agent to reason about and predict the actions of an object over time, rather than being concerned with the details of individual perceptions. The lowest level process in the figure simply processes low-level information to determine the presence of a single tap. This process has a buffer duration of 20 ms and a capacity of just a single element (in this case, the 'element' is actually a collection of low-level data). If more than one tap occurs within this time-frame, then the newer tap simply overwrites any previous tap. This is consistent with the experimental results that taps occurring within 20 ms of each other are merged in this way with the location of the newer tap dominating.

The next layer of processing consists of a single process with a buffer of duration 200 ms and a capacity of 2 elements. When a tap arrives at an empty buffer (which can happen at most once every 20 ms) a new aggregate conclusion is gener-

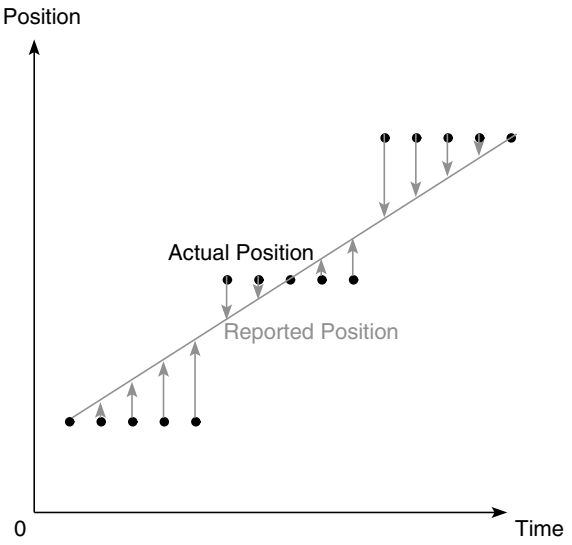


Figure 3: Actual vs. reported position of a series of taps. A train of taps is delivered with regular time interval to three positions on the arm. The subject reports that the locations of the taps are evenly distributed.

ated, taking the position of the tap as the start and end position of the 'run', and initialising the count of taps in this run to 1. This conclusion is passed to the output bus, where it is transmitted to other processes, but also, via a feedback loop, back to the input bus of the intermediate process. If a subsequent tap arrives before this aggregate fact expires from the buffer (i.e., within 200 ms) it overwrites the data about the previous tap (as this is older than the aggregate conclusion) and a new conclusion is formed which adjusts the end point of the run to the new tap position and increases the tap count by 1. The buffer duration ensures that any gap of 200 ms or more causes the previous 'run' to be forgotten, and thus any subsequent tap will be perceived as the start of a new run, which is consistent with the experimental data. It is important to note that although the buffer has a duration of 200 ms it is not necessary to delay conclusions for 200 ms. Instead, the process produces a conclusion whenever a new tap is felt (at most, once every 20 ms), and these conclusions can be acted upon immediately. For instance, Dennett [1991] suggests that a subject would be able to press a button after perceiving two taps at the wrist, but then still later report that the taps were evenly spaced. The process at the top-right of Figure 4 is waiting to do just that: it looks for a sequence of two taps at the wrist and then initiates the button press. The process on the left, however, is more conservative: it waits for a tap sequence to end before drawing a conclusion. (Detecting the end of a tap sequence can be done by employing a two-element capacity buffer and comparing the start position of sequential tap-run inputs). It is the output from this process that is eventually used to generate a report of the experience.

The model presented above suffices to explain the basic data of the cutaneous 'rabbit' experiments. Later work by Geldard [1977] appears appear to show that individual tap timings are preserved while the locations are displaced in inverse relation to the interval between the tap and a subsequent offset tap. An initial model of these results can be created by converting the count of taps in a run into a list of timings of individual taps. However, it is not clear if the results of these later experiments are entirely consistent with the original findings in [Geldard and Sherrick, 1972]. The initial findings that the train of taps is felt more or less evenly spaced would produce a situation as shown in Figure 3 where each tap location is adjusted towards an average. In the later experimental results, the adjustment is always towards a subsequent tap, which would suggest a report of bunching of taps towards the end of the train. While it is not clear at this stage which model is to be preferred, the architecture developed is capable of accommodating either.

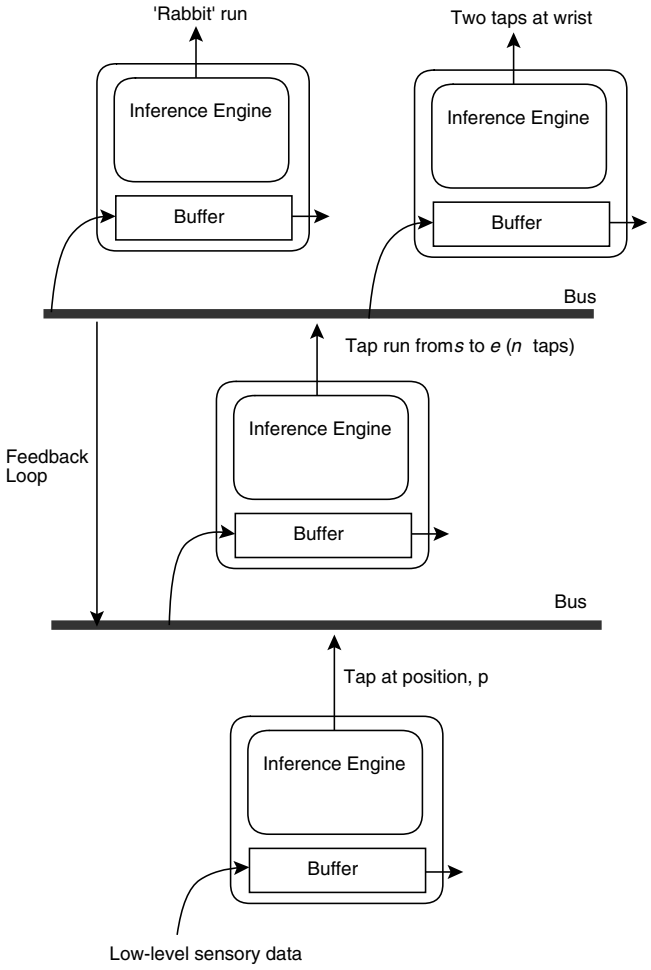


Figure 4: Network for cutaneous "rabbit" experiment.

5 Related Work

There has been considerable work on formal representations of time both by philosophers and Artificial Intelligence researchers (e.g., see [Allen, 1991] for a survey). However, there has been relatively little work on cognitive models of how humans represent and reason about events occurring over time at the reactive level. Most of the popular cognitive archi-

tectures such as Soar [Newell, 1990] and ACT-R [Anderson and Lebiere, 1998] concentrate on a primarily serial model of cognition which is less suited for modelling the sorts of experiments discussed in this paper. There has been some work done on modelling temporal perception in ACT-R [Taatgen *et al.*, 2004]. However this work concentrates on the perception of the passage of time itself, and uses a Temporal Module that can keep track of a single timer that can be used to measure and reproduce the interval between two events. In contrast, our work concentrates on the perception of phenomena in which the timing and sequence of events plays a crucial role. That is, we are studying the effects of time of perception rather than perception of time. In our architecture there are no explicit timers, or measurements of the passing of time; instead, the duration of buffers and the cycle time of inference processes can be used to make judgements about the temporal relationships between events.

The Multiple Drafts theory offers an alternative view of a highly parallel system with less of an overarching structure to cognition. However, to the best of our knowledge, there is no implemented cognitive architecture which captures all the features of the Multiple Drafts theory. Some parallel models of cognition come close to the "feel" of the Multiple Drafts theory. For example, the CopyCat architecture [Mitchell, 1993] is based on a Pandemonium-style collection of competing and cooperating stochastic processes. However, in contrast to the Multiple Drafts theory (and our architecture) which allows for both parallel processing and multiple simultaneous drafts, the CopyCat model employs many parallel processes working on a single solution (a single draft). The EPIC architecture [Kieras and Meyer, 1997] is also parallel, with multiple production rules executing simultaneously. However, unlike the Multiple Drafts theory, there is an explicit executive process in EPIC responsible for conflict resolution, resource arbitration and other *meta-management* tasks. The Global Workspace Theory [Baars, 1997] proposes a parallel, distributed architecture, but in contrast to the Multiple Drafts theory it explicitly advances the notion of a central workspace as a mechanism for sharing information and coordinating the parallel processes.

We are not aware of any previous attempts to model perceptual phenomena such as colour phi or saltation using these architectures. The architecture we have presented is capable of using separate processes to achieve the sort of integration found in the CopyCat, EPIC and Global Workspace models locally without requiring a single, global process, although it would be possible to construct a "central workspace" process within the architecture.

6 Discussion

The Multiple Drafts theory argues for replacing a central executive process in the brain with a parallel, decentralised view of processing. We agree with this view. A number of experiments, including the two discussed in this paper, suggest that a parallel view of human cognition is to be preferred, at least at the level of reactive perception. However, abandoning a *single* central executive process does not mean that information cannot be brought together *locally* for integration. The TAN architecture we have presented in this paper is a middle-road between serial central processing architectures and parallel Pandemonium architectures. The architecture is based on parallel networks of simple processes drawing conclusions based on individual snapshots of events occurring within a short time-span, connected via buses and feedback loops. TANs are capable of local (serial) integration while maintaining multiple simultaneous drafts: information flow can diverge as easily as converge. This conclusion is in contrast to that of Dennett and Kinsbourne who suggest that the only alternative to the Cartesian theatre is a strictly parallel architecture, where local integration is replaced by a more chaotic Pandemonium approach (e.g., [Kinsbourne, 1994] pp. 1324). The architecture and models we have presented allow for both separate analysis of aspects of a stimulus and local integration, without appealing to a central executive process where everything comes together.

The Temporal Abstraction Network models that we have developed demonstrate the ability of the architecture to model a variety of perceptual phenomena at the reactive level in which time and the temporal ordering of events plays a key role. In future work we plan to concentrate on extending the architecture to account for action selection, as well as expanding on the details of how reports are generated. One interesting area for future research will be to look at Libet's controversial experimental results [Libet, 1985] on voluntary action.

References

[Allen, 1991] James F. Allen. Time and time again: The many ways to represent time. *International Journal of Intelligent Systems*, 6(4):341–355, July 1991.

[Anderson and Lebiere, 1998] John R. Anderson and Christian Lebiere. *The Atomic Components of Thought*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 1998.

[Baars, 1997] Bernard J. Baars. In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness. *Journal of Consciousness Studies*, 4(4):292–309, 1997.

[Dennett and Kinsbourne, 1992] Daniel C. Dennett and Marcel Kinsbourne. Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences*, 15(2):183–247, 1992.

[Dennett, 1991] Daniel C. Dennett. *Consciousness Explained*. Penguin, London, 1991.

[Geldard and Sherrick, 1972] F. A. Geldard and C. E. Sherrick. The cutaneous 'rabbit': A perceptual illusion. *Science*, 178:178–179, 1972.

[Geldard, 1977] F. A. Geldard. Cutaneous stimulis, vibratory and saltatory. *Journal of Investigative Dermatology*, 69:83–87, 1977.

[Kieras and Meyer, 1997] David E. Kieras and David E. Meyer. An overview of the EPIC architecture for cognition and performance with application to human-computer interaction. *Human-Computer Interaction*, 12(4):391–438, 1997.

[Kinsbourne, 1994] Marcel Kinsbourne. Models of consciousness: Serial or parallel in the brain? In M. S. Gazzaniga, editor, *The Cognitive Neurosciences*, pages 1321–1330. MIT Press, Cambridge, MA, 1994.

[Kolers and von Grünau, 1976] P. A. Kolers and M. von Grünau. Shape and color in apparent motion. *Vision Research*, 16:329–335, 1976.

[Libet, 1985] Benjamin Libet. Unconscious cerebral initiative and the role of conscious will in voluntary action. *Behavioral and Brain Sciences*, 8:529–566, 1985.

[Logan, 2000] Brian Logan. A design study for the attention filter penetration architecture. In *Proceedings of the AISB'00 Symposium on How to design a functioning Mind*, pages 94–101, Birmingham, UK, April 2000. AISB, AISB.

[Mitchell, 1993] Melanie Mitchell. *Analogy-Making as Perception: a computer model*. MIT Press, Cambridge, MA, 1993.

[Newell, 1990] Allen Newell. *Unified Theories of Cognition*. Harvard University Press, 1990.

[Newell, 1992] Allen Newell. Précis of "Unified theories of cognition". *Behavioral and Brain Sciences*, 15:425–492, 1992.

[Norman and Long, 1996] Timothy J. Norman and Derek Long. Alarms: An implementation of motivated agency. In Mike Wooldridge, Joerg P. Müller, and Milind Tambe, editors, *Intelligent Agents II — Agent Theories, Architectures, and Languages (ATAL-95)*, volume 1037 of *LNAI*, pages 219–234. Springer-Verlag, 1996.

[SFN, 2005] Brain facts: a primer on the brain and nervous system, 2005.

[Sloman and Poli, 1996] Aaron Sloman and Riccardo Poli. SIM_AGENT: A toolkit for exploring agent designs. In Mike Wooldridge, Joerg Mueller, and Milind Tambe, editors, *Intelligent Agents Vol II (ATAL-95)*, pages 392–407. Springer-Verlag, 1996.

[Sloman, 1998] Aaron Sloman. Damasio, Descartes, alarms and meta-management. In *Proceedings IEEE International Conference Systems, Man, and Cybernetics (SMC 98)*, pages 2652–2657, Los Alamitos, California, 1998. IEEE, IEEE Computer Society Press.

[Taatgen *et al.*, 2004] Niels Taatgen, Hedderick van Rijn, and John Anderson. Time perception: Beyond simple interval estimation. In *Proceedings of the sixth International Conference on Cognitive Modeling*, pages 296–301, Pittsburgh, PA, 2004.

Prediction of the Behavioural Strategy in a Chemotaxis Search Task

Manuel A. Sánchez-Montañés
Universidad Autónoma de Madrid
Escuela Politécnica Superior
Campus de Cantoblanco, Madrid 28049, Spain
manuel.smontanes@uam.es

Tim C. Pearce
University of Leicester
Department of Engineering
Leicester LE1 7RH, UK
t.c.pearce@le.ac.uk

Abstract

In this paper we propose a theoretical framework based upon hidden Markov models for describing goal-oriented animal behaviours. These models are optimised in a virtual environment using a genetic algorithm, in order to assess the performance (such as energy efficiency or reward stability) of the naturally evolved behavioural strategy of the animal. In general the optimal solution does not depend trivially on the global properties of the environment, since the sensory information available locally to the animal is characterised by non-stationary statistics determined by the history of the animal's action selection. Here, we apply our framework to the problem of chemotactic search in the moth within naturally turbulent chemical plumes, and find that behavioural states emerge from the optimisation procedure which closely resemble those observed in real moth flight trajectories ("cast" and "surge"). Moreover, we find that the transition dynamics between these states also match qualitatively with biology. Thus, stereotypical chemotaxis behaviour in moths is correctly predicted by our framework, which we conclude provides an energy efficient solution to chemical source localisation.

1 Introduction

A common approach to understanding the behaviour of animals within their natural habitat involves breaking down complex behavioural sequences into simpler components termed *behavioural units* [Lenhner, 1998; Martin and Bateson, 1993; Harris and Foster, 1995]. Complex behaviours can be considered to consist of these behavioural units executed in some sequence to achieve an end goal. Important questions are how should these different behavioural units be characterised, and in what way should these be executed in order to optimise some overall performance criteria, such as energy minimisation, reward maximisation or stability. Answering these questions might account for qualitative and/or quantitative aspects of empirical data derived from behavioural experiments.

In the general case of an animal embedded within its environment these questions are complex, since the animal may change its own observed sensory statistics through its own actions, for example, by moving to another area or following a group of prey. This we term the *principle of locality*. Consequently, the optimal behavioural unit for the animal to execute at any given time can potentially depend upon the entire history of observed environmental states local to the observer, as well as the history of its interactions with that environment. Therefore any general theory of interaction between an agent and its environment must take into account the non-stationarity of sensory statistics, which are under some level of control by the agent through the sequence of behavioural primitives executed over time due to the principle of locality.

In this study we propose a theoretical framework which is based upon a stochastic model of agent-environment interaction that satisfies the principle of locality. Animal memory is represented here by having a variety of internal states. This contrasts with systems with no memory (one internal state) whose actions do not depend on the history of interactions with the environment. This arrangement is sufficient to satisfy the principle of locality, since the actions that the model performs depend not only on the internal state of the model, but also on the sensory input that it is currently sensing. The sensory input has the potential to change the internal state of the agent, whereas the actions change the state of the environment as well as the state of the agent with respect to that environment.

We will demonstrate that this framework allows us to predict and understand the structure of animal behaviours, by illustrating with a well characterised animal behaviour – chemotaxis in the moth – which is well characterised and performed robustly in turbulent chemical plumes. In this example we will see that the optimisation of a two-state model in a virtual plume generates a behavioural strategy which is directly comparable to that classically described in the animal.

2 Chemotactic search in turbulent plumes

The example we choose to illustrate our approach is moth chemotaxis in turbulent plumes (see Figure 1). Chemotaxis in moths is classically described by an alternation between two different behavioural states called *cast* and *surge* respectively [Baker, 1986; Vickers and Baker, 1994]. In the surge state the insect moves forward approximately against the direction of the oncoming wind, whereas in the cast state the insect oscillates perpendicularly to the wind with little forward

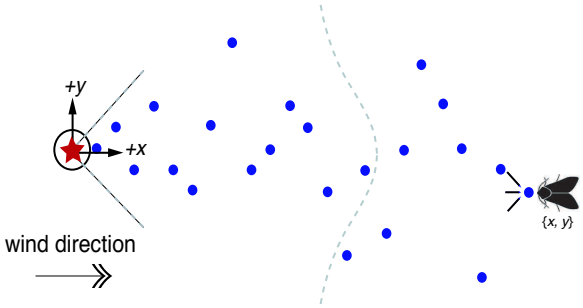


Figure 1: Model of the odour plume. The plume may be modeled as a number of discrete pockets, which move with fixed velocity away from the source (star at $\{x = 0, y = 0\}$) along the longitudinal axis (centreline) of the plume at the rate of one unit per time step, whilst dispersing away from the centreline as a random walk. At each time step an odour pocket moves one unit in the direction of the wind $(x + 1)$ and either $(y + 1)$, $(y - 1)$, or y, each with fixed probability $p = 1/3$.

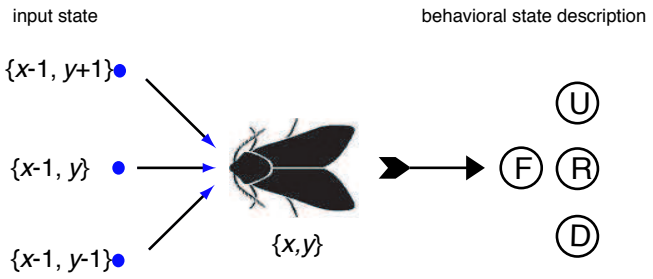


Figure 2: Moth basic actions. The moth has a specific $\{x, y\}$ location relative to the source at step n and is assumed to detect a pocket if a pocket moves into sector $\{x, y\}$ from sector $\{x - 1, y + 1\}$, $\{x - 1, y\}$ or $\{x - 1, y - 1\}$, otherwise no pocket is detected. At each step the moth must decide to execute one of four behavioural primitives based upon available sensory data (current input state and potentially its history): U - move up (to sector $\{x, y + 1\}$); F - move forward (to sector $\{x - 1, y\}$); D - move down (to sector $\{x, y - 1\}$); or R - rest (remain in sector $\{x, y\}$).

movement. Yet there are several questions which are not well understood in this context such as: How well adapted is the natural two-state solution to the problem ? Given that the behaviour is classically described by two behavioural units, how should each of them be defined in order to efficiently solve the task ? Finally, how should the transitions between them be determined by the external stimuli ?

In this context we consider sensorimotor rules that minimise the number of steps to locate the source (energy criterion). As in the general case, it is important to note that also in this context the statistics of the sensory input are not fixed - since at any given time step the sensory input depends upon *both* the state of the environment (in this case the plume structure) and the spatial position of the animal with respect to its environment. As such, an optimally efficient solution to the moth chemotaxis problem is likely to require some memory in the moth so that the history of its past interaction with the environment is taken into account in its future actions.

2.1 Model of the plume

We model the odour plume following [Balkovsky and Shraiman, 2002] (see Figure 1). The fluid dynamics of chemical plumes at behaviourally relevant flow velocities and viscosities generate high concentration "pockets" (filled circles), interspersed by air with almost no target odour [Murlis, 1986; Murlis *et al*., 1992; 2000]. The plume may be modeled as a number of discrete pockets, which move with fixed velocity away from the source (star at $\{x = 0, y = 0\}$) along the longitudinal axis (centreline) of the plume at the rate of one unit per step, whilst dispersing away from the centreline as a random walk (Figure 1). At each time step an odour pocket moves one unit in the direction of the wind $(x + 1)$ and either $y + 1$, $y - 1$, or y, each with fixed probability $p = 1/3$. This simple stochastic process generates a plume with pockets that are confined to an apex (lines starting at the star), but where the probability of finding a pocket in the y dimension is approximated by a Gaussian distribution with mean positioned at the centreline (dashed curve), and variance which depends upon the longitudinal distance from the source.

2.2 Moth basic actions

The moth has a specific $\{x, y\}$ location relative to the source at step n and is assumed to detect a pocket if a pocket moves into sector $\{x, y\}$ from sector $\{x - 1, y + 1\}$, $\{x - 1, y\}$ or $\{x - 1, y - 1\}$, otherwise no pocket is detected (Figure 2).

At each step the moth must decide to execute one of four behavioural primitives based upon available sensory data (current input state and potentially its history): U - move up (to sector $\{x, y+1\}$); F - move forward (to sector $\{x-1, y\}$); D - move down (to sector $\{x, y - 1\}$); or R - rest (remain in sector $\{x, y\}$). The simulation ends when either the moth moves into the sector containing the source, in which case a hit is recorded, or moves past the source ($x < 0$), in which case a miss is recorded. The problem is then to control the behavioural state transitions based upon the sensory input in order to minimise the number of steps required to locate the source, such that the hit rate is maximised.

3 Behavioural stochastic model

We model the dynamics of the animal as a hidden Markov model. It consists of a fixed number of internal states (in our case 1 or 2). In each one of them the probabilities $p(action|state)$ define the policy of the animal (which actions to perform in those states). Note that $action \in \{\text{Up}, \text{Down}, \text{Forward}, \text{Rest}\}$ and $state \in \{1, 2\}$. On the other hand, the transition dynamics between those internal states is defined by the probabilities $p(new\ state|old\ state,\ input)$ with the variable input being "no pocket", "pocket from y", "pocket from $y+1$" or "pocket from $y - 1$" (with y being the current moth vertical location), as explained in Section 2. Thus the two sets of probabilities $p(action|state)$ and $p(new\ state|old\ state,\ input)$

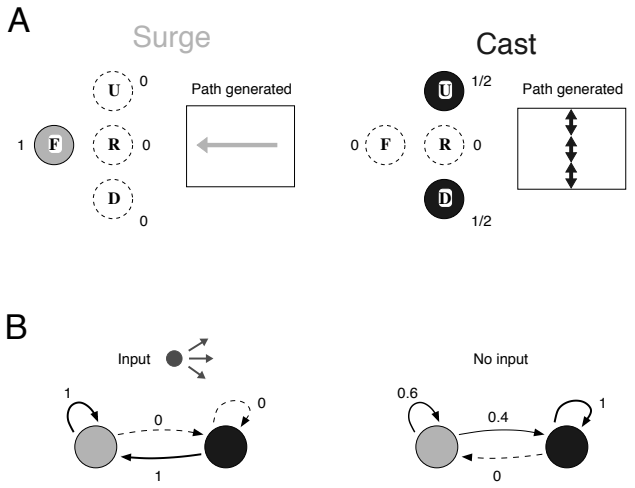


Figure 3: Optimal solution found by the genetic algorithm. **A**: Probability of performing the different actions in each of the two internal states. **B**: Transition probabilities between the internal states.

completely define the behavioural dynamics of the moth in our simulation, and they constitute the free parameters of our problem. We use a genetic algorithm [Levine, 1998] to find their optimal values which minimise the average time needed to reach the odour source. Thus, we calculate the optimal behavioural dynamics which minimises the time needed to find the source, whilst only imposing the total number of internal states and the behavioural primitives.

4 Results

When only one internal state is considered, there is no single strategy which efficiently finds the odour source (data not shown). However, for two internal states, there exists a strategy which finds the source with high probability. In Figure 3 we show the optimal solution found by the genetic algorithm in this case. Remember that the free parameters are the transition probabilities between the internal states (which we will call "grey" and "black"), and the probabilities of performing the different actions in those states.

When the Markov model is in the "grey" state the action performed is "Forward" with probability 1, so the model moves forward against the wind. In the "black" state the actions performed are "Up" and "Down" with probabilities equal to $1/2$, so the model oscillates perpendicularly to the wind. Thus the dynamics of these two very different states closely resemble the *surge* and *cast* behavioural units observed in behaving moths (grey and black respectively).

Let us analyze now what the found transition probabilities imply. When there is input from any direction, the moth changes to "surge" state with probability one. On the other hand, when there is no input the moth remains in the "cast" state in case it was already in it. In case the moth was in the "surge" state, it has some tendency to remain in it ($p = 0.6$), having a probability of 0.4 of changing to the "cast" state. These two probabilities determine how long the moth is on average in the "surge" state in the absence of input.

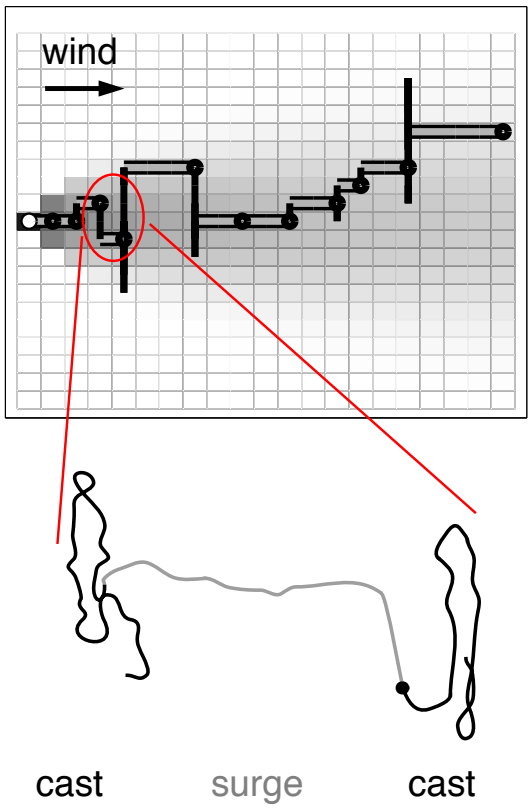


Figure 4: **Top:** Example of a path followed by the optimal solution. The path starts at the right circle and finally finds the source (white circle). The probability of finding an odour pocket is represented as the grey level of the pixel. Circles indicate locations at which the moth finds odour pockets in this particular simulation. The internal state of the moth at each stage of the path is represented by the path color: grey for "surge" state, and black for "cast" state. **Bottom:** real moth trajectory demonstrating cast/surge/cast behaviours. Adapted from [Mafra-Neto and Cardé, 1994].

In Figure 4 we show an example of a path followed by the optimal solution. This can be compared to real flight trajectories in the moth in response to chemical plumes (Figure 4, bottom), where contact with a single pheromone pulse was followed by a sharp turn upwind, and a faster and straighter upwind flight. In the absence of a second pheromone pulse, males returned to casting flight. All these features qualitatively match the optimal solution predicted by our framework.

5 Discussion

We have presented a framework that can be used to predict real behaviours for complex behavioural tasks, constrained a priori only by the total number of internal states and selection of behavioural primitives. . It consists of simulating the environment-animal interaction with the animal being mod-

eled as a Markov process. The parameters of this model are found through optimisation of a global behavioural cost function using a genetic algorithm.

We have illustrated this framework with a specific complex behaviour – moth chemotaxis. Our results show that when only one internal state is considered (which corresponds to a system with no memory), there is no single strategy which efficiently finds the odour source. However, for two internal states, there exists a strategy which finds the source with high probability, showing the importance of multiple states (memory) in this behavioural task.

It is clear how the optimal parameters of our solution relate to the behavioural strategy adopted by natural selection. In particular, the two internal states map qualitatively onto the observed behaviour of moths conducting chemotaxis - in that one state generates exploratory cross-wind search ("cast") behaviour and is preferred when no input has been detected, whilst the other state generates forward movement ("surge") behaviour which is preferred when input is detected. Thus our model suggests that, at least for dynamics with up to two internal states, the surge and cast behavioural units performed by the moth provide an efficient solution to chemical search in complex odour plumes. For our purposes, no further quantification of moth flight trajectories is required, since our aim is to explain the optimality of the chemotaxis strategy in the moth in terms of the two classical behavioural units identified by ethologists [Baker, 1986; Vickers and Baker, 1994].

We expect that the optimal parameters will depend on the physical properties of the environment - such as wind speed and dispersion in the plume. It would be interesting to predict the values of these parameters for different conditions.

The ultimate objective in building such models is to link behavioural requirements to solve a particular task efficiently with the corresponding mechanistic/architectural requirements of the behaviour and the underlying nervous system – for example, degree of memory, degree of behavioural complexity and even what degree of temporal coding of the sensory dynamics might be required.

Although illustrated with a behavioural case study, the framework outlined here is however general, in that it can be useful for defining the basic behavioural states needed for performing any complex behavioural task where the environmental dynamics are well defined. The model is compatible with the principle of locality in that the actions generated depend on both the state of the environment and the state of the model with respect to that environment. Thus, in this model the actions have direct consequences for future reinforcement and input from the environment which is directly relevant to the task solution. The framework we have described here can be applied to other well-defined behavioral sequences in complex environments with non-stationary statistics. Thus, we can extend this to address key questions related to natural action selection, such as

1. How complex does a behaviour need to be in order to efficiently solve a given task? In other words, what is the trade-off between number of behavioural units and task performance?
2. What is the memory requirement to solve a particular task? That is, to what degree is the past history of the sensory input and motor output crucial to solving the task efficiently?
3. How should the execution of specific behavioural units within a complex behaviour programme depend upon the sensory input? In other words, how should the transitions between behavioural units depend upon the sensory input?
4. How should different sensory modalities be integrated at different stages of the complex behaviour in order to efficiently solve the task?

Acknowledgments

The authors wish to acknowledge the support of the Royal Society (London) for awarding a Joint European Project grant that made this work possible. TCP was funded under the project AMOTH - A Fleet of Artificial Chemosensing Moths for Distributed Environmental Monitoring under the EU IST Future and Emerging Technologies programme (grant reference IST-2001-33066, project website http://www.amoth.org). MSM was also funded under the grant BFI2003-07276 from MCyT (Spain).

References

[Baker, 1986] T.C. Baker. Pheromone-modulated movements of flying moths. In T.L. Payne, M.C. Birch, and C.E.J. Kennedy, editors, *Mechanisms in Insect Olfaction*, pages 39–48, Oxford, UK, 1986. Oxford University Press.

[Balkovsky and Shraiman, 2002] E. Balkovsky and B.I. Shraiman. Olfactory search at high reynolds number. *PNAS*, 99(20):12589–93, 2002.

[Harris and Foster, 1995] M.O. Harris and S.P. Foster. Behavior and integration. In R.T. Carde and W.J. Bell, editors, *Chemical Ecology of Insects 2*, New York, 1995. Chapman & Hall.

[Lenhner, 1998] P.N. Lenhner. *Handbook of Ethological Methods*. Cambridge University Press, Cambridge, Massachusetts, 2nd edition, 1998.

[Levine, 1998] D. Levine. Pgapack parallel genetic algorithm library. *http://www.fp.mcs.anl.gov/CCST/research/reports_pre1998/comp_bio/stalk/pgapack.html*, 1998.

[Mafra-Neto and Cardé, 1994] A. Mafra-Neto and R.T. Cardé. Fine-scale structure of pheromone plumes modulates upwind orientation of flying moths. *Nature*, 369:142–4, 1994.

[Martin and Bateson, 1993] P. Martin and P. Bateson. *Measuring Behaviour*. Cambridge University Press, Cambridge, Massachusetts, 2nd edition, 1993.

[Murlis *et al*., 1992] J. Murlis, J.S. Elkinton, and R.T. Cardé. Odor plumes and how insects use them. *Annu. Rev. Entomol*., 37:505–32, 1992.

[Murlis *et al*., 2000] J. Murlis, M.A. Willis, and R.T. Cardé. Spatial and temporal structures of pheromone plumes in fields and forests. *Physiol. Entomol*., 25:211–22, 2000.

[Murlis, 1986] J. Murlis. The structure of odour plumes. In T. L. Payne, M. C. Birch, and C. E. J. Kennedy, editors, *Mechanisms in insect olfaction*, pages 27–38, Oxford, 1986. Clarendon Press.

[Vickers and Baker, 1994] N.J. Vickers and T.C. Baker. Reiterative responses to single strands of odor promote sustained upwind flight and odor source location by moths. *PNAS*, 9:5756–60, 1994.

Selecting Actions and Making Decisions: Lessons from AI Planning

Héctor Geffner
Departamento de Tecnología
ICREA – Universitat Pompeu Fabra
Barcelona 08003, SPAIN
hector.geffner@upf.edu

Abstract

Humans encounter a huge variety of problems which they must solve using general methods. Even simple problems, however, become computationally hard for general solvers if the structure of the problems is not recognized and exploited. Work in Artificial Intelligence Planning and Problem Solving has encountered a similar difficulty, leading in recent years to the development of well-founded and empirically tested techniques for recognizing and exploiting structure, focusing the search for solutions in certain cases, and bypassing the need to search in others. These techniques include the automatic derivation of heuristic functions, the use of limited but effective forms of inference, and the compilation of domains, all of which enable a general problem solver to 'adapt' automatically to the task at hand. In this paper, I present the ideas underlying these new techniques, and argue for their relevance to models of natural intelligent behavior as well. The paper is not a review of AI Planning – a diverse field with a long history – but a personal appraisal of some recent key developments and their potential bearing on accounts of action selection in humans and animals.

1 Introduction

In the late 50's, Newell and Simon introduced the first AI planner – the General Problem Solver or GPS – as a psychological theory [Newell and Simon, 1958; 1963]. Since then, Planning has remained a central area in AI while changing in significant ways: it has become more mathematical (a variety of planning problems has been clearly defined and studied) and more empirical (planners and benchmarks can be downloaded freely, and competitions are held every two years), and as a result, new ideas and techniques have been developed that enable the automatic solution of large and complex problems [Smith, 2003].

AI Planning studies languages, models, and algorithms for describing and solving problems that involve the selection of actions for achieving goals. In the simplest case, in *classical planning*, the actions are assumed deterministic, while in *contingent planning,* actions are non-deterministic and there is feedback. In all cases, the task of the planner is to compute a plan or solution; the *form* and *cost* of these solutions depending on the model; e.g., in classical planning, solutions are sequences of actions and cost is measured by the number of actions, while in planning with uncertainty and feedback, solutions map states into actions, and cost stands for expected or worst-possible cost.

Planning is a form of 'general problem solving' over a class of models, or more precisely, a *model-based* approach to intelligent behavior: given a problem in the form of a compact description of the actions, sensors (if any), and goals, a planner must compute a solution, and if required, a solution that minimizes costs. Some of the models used in planning, as for example Markov Decision Processes (MDPs), are not exclusive to AI Planning, and are used for example in Control Theory [Bertsekas, 1995], Reinforcement Learning [Sutton and Barto, 1998], and Behavioral Ecology [Houston and McNamara, 1988; Clark, 1991] among other fields. What is particular about AI planning are the *languages* for representing these models, the *techniques* for solving them, and the ways these techniques are *validated*. Techniques do matter quite a lot: even simple problems give rise to very large state spaces that cannot be solved by exhaustive methods. Consider the well known Rubik Cube puzzle: the number of possible configurations is in the order of trillions, yet methods are known for solving it, even optimally, from arbitrary configurations [Korf, 1998]. The key idea lies in the use of *admissible heuristic functions* that provide an optimistic approximation of the number of moves to solve the problem from arbitrary configurations. These functions enable the solution of large problems, even ensuring optimality, by focusing the search and avoiding most states in the problem. Interestingly, recent work in planning has shown that such functions can be derived *automatically* from the problem description [Bonet and Geffner, 2001], and can be used to drive the search in problems involving uncertainty and feedback as well [Bonet and Geffner, 2000]. Such functions can be understood as a specific and robust form of *means-ends analysis* [Newell and Simon, 1958; 1963] that produces goal-directed behavior in complex settings even in the presence of large state and action spaces.

In this paper, we review some of the key computational ideas that have emerged from recent work in planning and problem solving in AI, and argue that these ideas, although

not necessarely in their current form, are likely to be relevant for understanding natural intelligent behavior as well. Humans encounter indeed a huge variety of problems which they must solve using general methods. It cannot be otherwise, because there cannot be as many methods as problems. Yet, simple problems become computationally hard for a general solver if the structure of the problems is not recognized and exploited. This is well known in AI, where systems that do not exhibit this ablity tend to be shallow and brittle. In the last few years, however, work in Planning and Problem Solving has led to well-founded and empirically tested techniques for recognizing and exploiting structure, focusing the search for solutions, and in certain cases, bypassing the need to search altogether. These techniques include the automatic derivation of heuristic functions, the use of limited but effective forms of inference, and the compilation of domains, all of which enable a general problem solver to 'adapt' automatically to the task at hand. Interestingly, the need for focusing the search for solutions has been recognized in a number of recent works concerned with natural intelligent behavior, where it has been related to the role of emotions in the appraisal and solution of problems. We will say more about this as well.

Since Newell's and Simon's GPS, the area of AI planning has departed from the original motivation of understanding human cognition to become the mathematical and computational study of the problem of selecting actions for achieving goals. Yet after all these years, and given the progress achieved, it is time to reflect on what has been learned in the abstract setting, and use it for informing our theories in the natural setting. This exercise is possible and may be quite rewarding. It parallels the approach advocated by David Marr, and echoed more recently by [Glimcher, 2003] and others in the Brain Sciences; namely: characterize *what* needs to be computed, *how* it can be computed, and how these computations are *approximated* in real-brains. The findings that we summarize below, aim to provide a partial account of the first two tasks.

A few methodological comments before proceeding. First about *domain-generality* vs. *domain-specificity* in action selection. I have said that humans are capable of solving a wide range of problems using general methods. This, however, is controversial. Both evolutionary psychologists [Tooby and Cosmides, 1992] and cognitive scientists from the 'fast and frugal heuristics' school [Gigerenzer and Todd, 1999] place an emphasis on modularity and domain-specificity. Others, without necessarily denying the role of specialization, postulate the presence of general reasoning and problem solving mechanisms as well, at least in humans (see for example [Stanovich, 2004]). We are not going to address this controversy here, just emphasize that 'general' and 'adapted' are not necessarily opposite of each other. Indeed, the work in AI planning is domain-independent, yet the recent techniques illustrate how a general problem solver can 'adapt' to a specific problem by recognizing and exploiting structure, for example, in the form of heuristic functions. These heuristics are indeed in line with the 'fast and frugal heuristics', the difference being that they are general and can be extracted automatically from problem descriptions.

Another distinction that is relevant for fitting the work in AI Planning within the broader work on Intelligent Behavior is the one between *finding solutions* vs. *executing solutions*. For many models, such as those involving uncertainty and feedback, the solutions, from a mathematical point of view, are functions mapping states into actions (these functions are called *closed-loop policies*, and in the partially observable case map actually *belief states* into actions; see below). These functions can be represented in many ways; e.g. as condition, action rules, as value functions, etc. Indeed, in what is often called *behavior-based AI* [Brooks, 1997], these solutions are encoded by hand for controlling mobile robots. In nature, similar solutions are thought to be encoded in brains but not by hand but by evolution. Representing and executing solutions, however, while challenging, is different than coming up with the solutions in the first place which is what AI Planning is all about. Whether this is a requirement of intelligent behavior in animals is not clear although it seems to be a distinctive feature of intelligent behavior in humans. Interestingly, in many cases, the same models can be used for both *understanding* the solutions found in nature, and for *generating* those solutions [McFarland and Bosser, 1993]. The interest in the latter case, however, is not only with the models but also with the algorithms needed for solving those models effectively. We thus consider both models and algorithms.

2 Models

Most models considered in AI Planning can be understood in terms of *actions* that affect the *state* of a system, and can be given in terms of

1. a discrete and finite state space S,
2. an initial state $s_0 \in S$,
3. a non-empty set of terminal states $S_T \subseteq S$,
4. actions $A(s) \subseteq A$ applicable in each non-terminal state,
5. a function $F(a, s)$ mapping non-terminal states s and actions $a \in A(s)$ into *sets* of states
6. action costs $c(a, s)$ for non-terminal states s, and
7. terminal costs $c_T(s)$ for terminal states.

In deterministic planning, there is a single predictable next state and hence $|F(a, s)| = 1$, while in non-neterministic planning $|F(a, s)| \geq 1$. In addition, in probabilistic planning (MDPs), non-deterministic transitions are weighted with probabilities $P_a(s'|s)$ so that $\sum_{s' \in F(a,s)} P_a(s'|s) = 1$. In general, action costs $c(a, s)$ are assumed to be positive, and terminal costs $c_T(s)$ non-negative. When zero, terminal states are called *goals*. The models underlying 2-player games such as Chess can be understood also in these terms with opponent moves modeled as non-deterministic transitions. Often models are described in terms of rewards rather than costs, or in terms of both, yet care needs to be taken so that models have well-defined solutions. State models of this type are also considered in Control Theory [Bertsekas, 1995], Reinforcement Learning [Sutton and Barto, 1998], and Behavioral Ecology [Houston and McNamara, 1988; Clark, 1991]. In [Astrom, 1965], it is shown how problems involving partial feedback can be reformulated as problems involving full state feedback over *belief states*; i.e., states that

encode the information about the true state of the system. All these problems can also be cast as *search problems* in either the original state space or belief space [Bonet and Geffner, 2000].

The solutions to these various state models have a mathematical form that depends on the type of feedback. In problems without feedback, solutions are sequences of actions, while in problems with full-state feedback solutions are functions mapping states into actions (called also closed-loop control policies). The form of the solution to the various models need to be distinguished from the way they are represented. A common, compact representation of policies is in terms of condition, action rules; yet many of the standard algorithms assume a representation of policies in terms of less-compact value functions. The problem of combining robust algorithms with compact representations is not yet solved, although significant progress has been achieved when actions can be assumed to be deterministic.

From a complexity point of view, if there are n variables, the state space (range of possible value assignments) is exponential in n. Thus, except for problems involving very few variables, exhaustive approaches for specifying or solving these models are unfeasible. A key characteristic of AI Planning are the languages for representing these models, and the techniques used for solving them.

3 Languages

A standard language for representing state models in compact form is Strips [Fikes and Nilsson, 1971].[1] In Strips, a problem P is expressed as a tuple $P = \langle A, O, I, G\rangle$ where A is the set of atoms or boolean variables of interest, O is the set of actions, and $I \subseteq A$, and $G \subseteq A$ are the atoms that are true in the initial and goal situations respectively. In addition, each action $a \in O$ is characterized by three sets of atoms: the atoms $pre(a)$ that must be true in order for the action to be executable (preconditions), the atoms $add(a)$ that become true after the action is done (add list), and finally, the atoms $del(a)$ that become false after doing the action (delete list).

A Strips planner is a *general problem solver* that accepts descriptions of arbitrary problems in Strips, and computes a solution for them; namely, sequences of actions mapping the initial situation into the goal. Actually, any deterministic state model can be expressed in Strips, and any Strips problem $P = \langle A, O, I, G\rangle$ defines a precise state model $S(P)$ where

- the states s are the different subsets of atoms in A
- the initial state s_0 is I
- the goal states s_G are those for which $G \subseteq s_G$
- $A(s)$ is the subset of actions $a \in O$ s.t. $pre(a) \subseteq s$
- $F(a, s) = \{s + Add(a) - Del(a)\}$, for $a \in A(s)$
- the actions costs $c(a, s)$ are uniform (e.g., 1)

Extensions of the Strips language for accommodating non-boolean variables and other features have been developed, and planners capable of solving large and complex problems currently exist. This is the result of new ideas and a solid empirical methodology in AI Planning following [Penberthy and Weld, 1992], [Blum and Furst, 1995], and others in the 90's.

4 Is Strips Planning relevant at all?

Before getting into the techniques that made this progress possible, let us address some common misconceptions about Strips planning. First, it is often said that Strips planning cannot deal with uncertainty. This is true in one way, but not in another. Namely, the model $S(P)$ implicit in a Strips encoding P does not *represent* uncertainty. Yet this does not imply that Strips planning cannot *deal* with uncertainty. It actually can. Indeed, the 'winner' of the ICAPS 2004 Probabistic Planning Competition [Littman, 2005], FF-Replan,[2] is based on a Strips planner called FF [Hoffmann and Nebel, 2001]. While the actions in the domain were probabilistic, FF-Replan ignores the probabilities and replans from scratch using FF after every step. Since currently, this can be done extremely fast even in domains with hundred of actions and variables, this deterministic re-planner did better than more sophisticated probabilistic planners. It does not take much to see that this strategy may work well in a 'noisy' Block Worlds domain where blocks may accidentally fall off gripper, and actually it is not trivial to come up with domains where this strategy will not work (this was indeed the problem in the competition). Control engineers know this very well: stochastic systems are often controlled by closed-loop control policies designed under deterministic approximations, as in many cases errors in the model can be safely corrected through the feedback loop.

A second misconception about Strips or 'classical' planning is that actions denote 'primitive operations' that all take a unit of time. This is not so: Strips planning is about planning with operators that can be chacterized in terms of pre and postconditions. The operator themselves can be abstractions of lower level policies, dealing with low level actions and sensors. For example, the action of grabbing a cup involves moving the arm in certain ways, sensing it, and so on; yet for higher levels, it is natural to assume that the action can be summarized in terms of preconditions involving the proximity of the cup, a free-hand, etc; and postconditions involving the cup in the hand and so on. Reinforcement learning has been shown to be a powerful approach for learning low-level skills, but it has been less successful for integrating these skills for achieving high-level goals. The computational success of Strips planning suggests that one way of doing this is by characterizing low-level behaviors in terms of pre and postconditions, and feeding such behaviors into a planner.

5 Heuristic Search

How can current Strips planners assemble dynamically and effectively low-level behaviors, expressed in terms of pre and post conditions, for achieving goals? The idea is simple: they exploit the structure of the problems by extracting automatically informative heuristic functions. While the idea of using

[1] Strips is the name of a planner developed in the late 60's at SRI, a successor of Newell's and Simon's GPS.

[2] FF-Replan was developed by SungWook Yoon, Alan Fern and Robert Givan from Purdue.

heuristic functions for guiding the search is old [Hart *et al*., 1968], the idea of extracting these functions automatically from problem encodings is more recent [McDermott, 1996; Bonet *et al*., 1997], and underlies most current planners.

In order to illustrate the power of heuristic functions for guiding the search, consider the problem of looking in a map for the shortest route between Los Angeles and New York. One of the best known algorithms for finding shortest routes is Dijkstra's algorithm [Cormen *et al*., 1989]: the algorithm efficiently and recursively computes the shortest distances $g(s)$ between the origin and the closest 'unvisited' cities s til the target is reached. A characteristic of the algorithm when applied to our problem, is that it would first find a shortest path from LA to Mexico City, even if Mexico City is way out of the best path from LA to NY. Of course, this is not the way people find routes in a map. The *heuristic search algorithms* developed in AI approach this problem in a different way, taking into account an estimate $h(s)$ of the cost (distance) to go from s to the goal. In route finding, this estimate is given by the Euclidian distance in the map that separates s from the goal. Using then the sum of the cost $g(s)$ to get to s and an estimate $h(s)$ of the cost-to-go from s to the goal, heuristic or informed search algorithms are much more *focused* than blind search algorithms like Dijkstra, without sacrificing optimality. For example, in finding a route from Los Angeles to New York, heuristic search algorithms like A* or IDA* [Pearl, 1983; Russell and Norvig, 1994], would never consider 'cities' whose value $g(s) + h(s)$ is above the cost of the problem. These algorithms guarantee also that the solutions found are optimal provided that the heuristic function h is *admissible* or *optimistic*, i.e., if for any s, $h(s) \leq V^*(s)$, where V^* is the optimal cost function. In the most informed case, when $h = V^*$, heuristic search algorithms are completely focused and consider only states along optimal paths, while in the other extreme, if $h = 0$, they consider as many states as Dijkstra's algorithm. Most often, we are not in either extreme, yet good informed heuristics can be found that reduce the space to search quite drastically. For example, while with today's technology it is possible to explore in the order of 10^{10} states, optimal solutions to arbitrary configuration of the Rubik's Cube with more than 10^{20} states, have been reported [Korf, 1998]. These search methods are very selective and consider a tiny fraction of the state space only, smaller actually than $1/10^{10}$.

6 Deriving Heuristic Functions

Two key questions arise: 1) How can these heuristics be obtained? and 2) Whether similar gains can be obtained in other models, e.g., when actions are not deterministic and states are not necessarily fully observable. We address each question in turn.

The power of current planners arises from methods for extracting heuristic values $h(s)$ automatically from problem encodings. The idea is to set the estimated costs $h(s)$ of reaching the goal from s to the cost of solving a simpler, relaxed problem. Strips problems, for example, can be relaxed by dropping the delete lists. Solving (non-optimally) a delete-free Strips problem can be done quite efficiently, and the heuristic $h(s)$ can be set to the cost of the relaxation. The idea of obtaining heuristics by solving relaxed problems is old [Pearl, 1983], but the use of Strips relaxations for deriving them automatically for planning is more recent [McDermott, 1996; Bonet *et al*., 1997]. Since then other relaxations have been been considered. In [Bonet *et al*., 1997], the derived heuristics are used for selecting actions greedily, in real-time, without finding a complete plan first. The proposal is closely related to the *spreading activation model* of action selection in [Maes, 1990], with *activation levels* replaced by or interpreted as *heuristic values* (cost estimators).

The automatic derivation of heuristic functions for guiding the search provides what is probably the first fast and robust mechanism for carring out means-end analysis in complex domains.

7 Greedy Selection and Lookahead

Heuristic functions, as cost estimators, have also been found crucial for focusing the search in problems involving uncertainty and feedback where solutions are not 'paths' in the state space. Solutions to the various models can be all expressed in terms of control policies π that are *greedy* with respect to a given heuristic function h. A control policy π is a function mapping states $s \in S$ into actions $a \in A(s)$, and a policy π_h is greedy with respect to h iff π_h is the best policy assuming that the cost-to-go is given by h, i.e.

$$\pi_h(s) \;=\; \operatorname*{argmin}_{a \in A(s)} Q_h(a, s) \qquad (1)$$

where $Q_h(a, s)$ is the expression of the cost-to-go whose actual form depends on the model; e.g., for non-deterministic models is $c(a, s) + \max_{s' \in F(a,s)} h(s')$, for MDPs $c(a, s) + \sum_{s' \in F(a,s)} P_a(s'|s)h(s')$, etc. In all cases, if the heuristic h is optimal; i.e., $h = V^*$, the greedy policy π_h is optimal as well [Bellman, 1957; Bertsekas, 1995]. As mentioned above, the planner that won that the last Probabilistic Planning Competition, used a greedy policy based on an heuristic function derived ignoring probabilistic information.

Often, if the heuristic estimator h is good, the greedy policy π_h based on it is good as well. Otherwise, there are two ways for improving the policy π_h without having to consider the entire state space: one is by *look ahead*, the other is by *learning,* and both involve *search*. Look-ahead is the strategy used in 2-player games like Chess that cannot be solved up to the terminal states; it is a variation of the greedy strategy π_h where the $Q_h(a, s)$ term is obtained not from the direct successors of s but from further descendants. The lookahead search is not exhaustive either, as values $h(s')$ of the tip nodes are used to prune the set of nodes considered; a technique known as alpha-beta search [Newell *et al*., 1963]. The quality of the play depends on the search horizon and on the quality of the value function, which in this case, does not estimate cost but reward. In all the models, the greedy policy π_h is invariant to certain types of transformation in h; e.g $\pi_h = \pi_{h'}$ if $h' = \alpha h + \beta$ for constants α and β, $\alpha > 0$, so the value scale is not critical. Moreover, in Chess, any transformation of the heuristic function that preserves the relative ordering

of the states, yields an equivalent policy, even if lookahead is used.

8 Learning

The second way to improve a greedy policy π_h is by adjusting the heuristic values during the search [Korf, 1990; Barto *et al*., 1995]. More precisely, after applying the greedy action $\text{argmin}_{a \in A(s)} Q_h(a, s)$ in state s, the heuristic value $h(s)$ in s is updated to

$$h(s) := \min_{a \in A(s)} Q_h(a, s) \quad (2)$$

Interestingly, if h is admissible ($h \leq V^*$), and these updates are performed as the greedy policy π_h is simulated, the resulting algorithm exhibits two properties that distinguish it from standard methods: first, unlike a fixed greedy policy, it will never get trapped into a loop and will eventually get to the goal (if the goal is reachable from every state), and second, after repeated trials, the greedy policy π_h converges to an optimal policy, and the values $h(s)$ to the optimal values $V^*(s)$ (over the relevant states). This algorithm is called Real-Time Dynamic Programming (RTDP) in [Barto *et al*., 1995] as it combines a greedy, real-time action selecting mechanism, with the improvements brought about by the updates. Like heuristic search algorithms in AI but unlike standard DP methods, RTDP can solve large problems involving uncertainty, without having to consider the whole state space, provided that a good and admissible heuristic function h is used. Moreover, partial feedback can be accommodated as well, by performing the search in 'belief space'. GPT is a planner, that accepts description of problems involving stochastic actions and sensors, and computes optimal or approximate optimal policies using a refinement of these methods [Bonet and Geffner, 2000; 2003].

9 Inference

Many problems have a low polynomial complexity, and are easy for people to solve; e.g., the problem of collecting packages at various destinations in a city, and delivering them at some other destinations. This 'problem' is not even considered a problem by people as, unlike puzzles, can be solved (non-optimally) in a very straightforward way. Yet if the problem is fed to a planner by describing the actions of driving the truck from one location to another, picking up and loading the packages, and so on, the planner would tackle the problem in the same way it would tackle a puzzle: by means of *search*. This search can often be done quite fast, yet like in Chess, this does not seem to be the way people solve such problems. Psychologists interested in problem solving, have focused on puzzles like Towers-of-Hanoi rather than on the simple problems that people solve every day. The work in planning however reveals that problems that are easy for people are not necessarily easy for a general automated problem solver. It may be argued that people solve these problems by using domain-specific knowledge, yet this pushes the problem one level up: how do people recognize when a problem falls in a domain, and how many domains are there? Recently we have addressed the related question of whether it is possible to solve a wide variety of 'simple' problems that are used as benchmarks in planning (including the famous Blocks World problems), by performing efficient (low polynomial) inference and *no search*. To our surprise, we have found that this is possible [Vidal and Geffner, 2005]. We believe that there are a number of useful consequences to draw from this fact, given that most problems faced by real intelligent agents are not puzzles. In any case, inference and heuristic functions are two sides of the same coin: they both extract useful knowledge from a domain description and use it to focus the search, and if possible, to eliminate the search altogether.

10 SAT: Search and Inference in Logic

Logic has played a prominent role in AI as a basis for knowledge representation and programming languages. In recent years, logic restricted to propositional languages has become a powerful computational paradigm as well. A variety of problems can be encoded as SAT problems which are then fed and solved by powerful SAT solvers: programs that take a set of clauses (disjunctions of positive or negated atoms), and determine if the clauses are consistent, and if so, return a truth-valuation that satisfies all the clauses (a model). While the SAT problem is intractable, problems involving thousands of clauses and variables can now be solved [Kautz and Selman, 2005]. Classical planning problems can be mapped into SAT by translating the problem descriptions into propositional logic, and fixing a planning horizon: if the theory is inconsistent, the problem has no solution within the horizon, else a plan can be read off the model [Kautz and Selman, 1996]. For problems involving non-determinism, the SAT formulation yields only 'optimistic' plans, yet work is underway for reproducing the practical success of SAT in richer settings. Current SAT algorithms combine search and inference as well, and are complete. Some of the original algorithms, were based on local search [Selman *et al*., 1992], and were inspired by a neural-network constraint satisfaction engine [Adorf and Johnston, 1990].

11 Domain Compilation

Another recent development in logic relevant for action selection is *knowledge compilation* [Selman and Kautz, 1996; Darwiche and Marquis, 2002b]. In knowledge compilation, a formula is mapped into a logically equivalent formula of a certain form that makes certain class of operations more efficient. For example, while testing consistency of a formula is exponential in the size of the formula (in the worst case), formulas in d-DNNF can be tested in linear time (d-DNNF is a variation of 'Disjunctive Normal Form'; see [Darwiche, 2001; 2002]). Moreover, for a formula T in d-DNNF, it is possible, in linear-time as well (i.e., very efficiently) to check the consistency of $T + L$ for any set of literals L, get a model of $T + L$, and even count the number of such models. Of course, compiling a formula into d-DNNF is expensive, but this expense is worth if the result of the compilation is used many times. The idea of theory compilation has a number of applications in planning that are beginning to get explored. For example, Barret in [Barret, 2004], compiles planning theories with a fixed planning horizon n

into d-DNNF, and shows that from the compiled theory *it is possible to obtain plans for arbitrary initial situations and goals, in linear-time with no search*. This is a very interesting idea that makes technical sense of the intuition that there are many logically equivalent representations, and yet some representations that are better adapted for a given task. We are currently exploring a variation of Barret's idea that exploits another property of d-DNNF formulas T: the ability to efficiently compute not only models of T but also *best* models, 'best' defined in terms of a ranking over the individual (boolean) variables [Darwiche and Marquis, 2002a]. By ranking the literals at the horizon n and then using ideas similar to ones above, it is possible to get in-linear time, for any initial situation, the best plan and the rank of the (best) situation that it leads to. In this way, very quickly, *we get an appraisal and appropriate 'reaction' to any situation, without doing any search*. It does not take much to relate these appraisals with the role *emotions* in the selection of actions as postulated in a number of recent works; e.g., [Damasio, 1995; Ketelaar and Todd, 2001; Belavkin, 2001; Evans, 2002; MacLeod, 2002]. In the view that arises from domain compilation, however, emotions are prior to search, and they are not used for guiding the search or deliberation, nor are they the result of deliberation; rather they summarize expected reward or penalty (as when a deer sees a lion nearby). This view can account also for the way in which local preferences (ranks) are quickly assembled to provide an assessment of any situation (see "Feeling and Thinking: Preferences Need No Inferences" in [Zajonc, 2004]). Computationally, the account is limited in two ways: it assumes a given fixed planning horizon, and that there is no uncertainty. Still it appears as a good starting point. In relation to the heuristic view of emotions, the notion of domain compilation provides an alternative and probably complementary view: in one case, emotion like an heuristic, guides the search for best reward, in the other, emotion stands for expected reward, which in the compiled representation is computed in linear-time (i.e., very quickly).

12 Summary

We have argued that humans encounter a huge variety of problems which they must solve using general methods. For general methods to work, however, they must be able to recognize and exploit structure. We have then reviewed some recent techniques from AI planning and problem solving that accomplish this, either focusing the search for solutions or bypassing the search altogether. These techniques include the automatic derivation of heuristic functions, the use of limited but effective forms of inference, and the compilation of domains, all of which enable a general problem solver to 'adapt' to the task at hand. We have also discussed briefly how these ideas relate to some biologically-motivated action selection models based on 'activation levels' and recent proposals linking emotions and search.

The area of planning and problem solving in Artificial Intelligence has come a long way, and it is probably time, following Marr's approach, to use the insights gained by the study of *what* is to be computed and *how* is to be computed, for gaining a better understanding of what real-brains actually compute when making plans and selecting actions. Of course, there is a lot to be learned, and many other useful and necessary approaches to the problem, yet some of us hope that a good theory of AI planning and problem solving, as Newell, Simon, and others envisioned many years ago, will be an essential part of the global picture.

References

[Adorf and Johnston, 1990] H. M. Adorf and M. D. Johnston. A discrete stochastic neural network algorithm for constraint satisfaction problems. In *Proc. the Int. Joint Conf. on Neural Networks*, 1990.

[Astrom, 1965] K. Astrom. Optimal control of markov decision processes with incomplete state estimation. *J. Math. Anal. Appl.*, 10:174–205, 1965.

[Barret, 2004] T. Barret. From hybrid systems to universal plans via domain compilation. In *Proc. ICAPS-04*, 2004.

[Barto *et al.*, 1995] A. Barto, S. Bradtke, and S. Singh. Learning to act using real-time dynamic programming. *Artificial Intelligence*, 72:81–138, 1995.

[Belavkin, 2001] R. V. Belavkin. The role of emotion in problem solving. In *Proc. of the AISB'01 Symposium on Emotion, Cognition and Affective Computing*, pages 49–57, 2001.

[Bellman, 1957] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[Bertsekas, 1995] D. Bertsekas. *Dynamic Programming and Optimal Control, Vols 1 and 2*. Athena Scientific, 1995.

[Blum and Furst, 1995] A. Blum and M. Furst. Fast planning through planning graph analysis. In *Proceedings of IJCAI-95*, pages 1636–1642. Morgan Kaufmann, 1995.

[Bonet and Geffner, 2000] B. Bonet and H. Geffner. Planning with incomplete information as heuristic search in belief space. In *Proc. of AIPS-2000*, pages 52–61. AAAI Press, 2000.

[Bonet and Geffner, 2001] B. Bonet and H. Geffner. Planning as heuristic search. *Artificial Intelligence*, 129(1–2):5–33, 2001.

[Bonet and Geffner, 2003] B. Bonet and H. Geffner. Labeled RTDP: Improving the convergence of real-time dynamic programming. In *Proc. 13th Int. Conf. on Automated Planning and Scheduling (ICAPS-2003)*, pages 12–31. AAAI Press, 2003.

[Bonet *et al.*, 1997] B. Bonet, G. Loerincs, and H. Geffner. A robust and fast action selection mechanism for planning. In *Proceedings of AAAI-97*, pages 714–719. MIT Press, 1997.

[Brooks, 1997] R. Brooks. From earwigs to humans. *Robotics and Autonomous Systems*, 20(2-4):291–304, 1997.

[Clark, 1991] C. Clark. Modeling behavioral adaptations. *Behavioral and Brain Sciences*, 14(1), 1991.

[Cormen *et al*., 1989] T. H. Cormen, C. E. Leiserson, and R. L. Rivest. *Introduction to Algorithms*. The MIT Press, 1989.

[Damasio, 1995] A. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Quill, 1995.

[Darwiche and Marquis, 2002a] A. Darwiche and P. Marquis. Compilation of propositional weighted bases. In *Proc. NMR-02*, 2002.

[Darwiche and Marquis, 2002b] A. Darwiche and P. Marquis. A knowledge compilation map. *J. of AI Research*, 17:229–264, 2002.

[Darwiche, 2001] Adnan Darwiche. Decomposable negation normal form. *J. ACM*, 48(4):608–647, 2001.

[Darwiche, 2002] A. Darwiche. On the tractable counting of theory models and its applications to belief revision and truth maintenance. *J. of Applied Non-Classical Logics*, 2002.

[Evans, 2002] D. Evans. The search hypothesis of emotion. *British J. Phil. Science*, 53:497–509, 2002.

[Fikes and Nilsson, 1971] R. Fikes and N. Nilsson. STRIPS: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 1:27–120, 1971.

[Gigerenzer and Todd, 1999] G. Gigerenzer and P. Todd. *Simple Heuristics that Make Us Smart*. Oxford, 1999.

[Glimcher, 2003] P. Glimcher. *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics*. MIT Press, 2003.

[Hart *et al*., 1968] P. Hart, N. Nilsson, and B. Raphael. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. Syst. Sci. Cybern*., 4:100–107, 1968.

[Hoffmann and Nebel, 2001] J. Hoffmann and B. Nebel. The FF planning system: Fast plan generation through heuristic search. *Journal of Artificial Intelligence Research*, 14:253–302, 2001.

[Houston and McNamara, 1988] A. I. Houston and J. M. McNamara. A framework for the functional analysis of behaviour. *Behavioral and Brain Sciences*, 11(1), 1988.

[Kautz and Selman, 1996] H. Kautz and B. Selman. Pushing the envelope: Planning, propositional logic, and stochastic search. In *Proceedings of AAAI-96*, pages 1194–1201. AAAI Press / MIT Press, 1996.

[Kautz and Selman, 2005] H. Kautz and B. Selman. The state of SAT. *Discrete and Applied Math*, 2005.

[Ketelaar and Todd, 2001] T. Ketelaar and P. M. Todd. Framing our thoughts: Evolutionary psychology's answer to the computational mind's dilemma. In H.R. Holcomb III, editor, *Conceptual Challenges in Evolutionary Psychology*. Kluwer, 2001.

[Korf, 1990] R. Korf. Real-time heuristic search. *Artificial Intelligence*, 42:189–211, 1990.

[Korf, 1998] R. Korf. Finding optimal solutions to Rubik's cube using pattern databases. In *Proceedings of AAAI-98*, pages 1202–1207. AAAI Press / MIT Press, 1998.

[Littman, 2005] Michael Littman. The first probabilistic planning competition. To be published, 2005.

[MacLeod, 2002] W. Bentley MacLeod. Complexity, bounded rationality and heuristic search. *Contributions to Economic Analysis & Policy*, 1(1), 2002.

[Maes, 1990] P. Maes. Situated agents can have goals. *Robotics and Autonomous Systems*, 6:49–70, 1990.

[McDermott, 1996] D. McDermott. A heuristic estimator for means-ends analysis in planning. In *Proc. Third Int. Conf. on AI Planning Systems (AIPS-96)*, 1996.

[McFarland and Bosser, 1993] David McFarland and Thomas Bosser. *Intelligent behaviour in animals and robots*. MIT Press, 1993.

[Newell and Simon, 1958] A. Newell and H. Simon. Elements of a theory of human problem solving. *Psychology Review*, 1958.

[Newell and Simon, 1963] A. Newell and H. Simon. GPS: a program that simulates human thought. In E. Feigenbaum and J. Feldman, editors, *Computers and Thought*, pages 279–293. McGraw Hill, 1963.

[Newell *et al*., 1963] A. Newell, J. C. Shaw, and H. Simon. Chess-playing programs and the problem of complexity. In E. Feigenbaum and J. Feldman, editors, *Computers and Thought*, pages 109–133. McGraw Hill, 1963.

[Pearl, 1983] J. Pearl. *Heuristics*. Addison Wesley, 1983.

[Penberthy and Weld, 1992] J. Penberthy and D. Weld. Ucpop: A sound, complete, partiall order planner for adl. In *Proceedings KR'92*, 1992.

[Russell and Norvig, 1994] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1994.

[Selman and Kautz, 1996] Bart Selman and Henry Kautz. Knowledge compilation and theory approximation. *Journal of the ACM*, 43(2):193–224, 1996.

[Selman *et al*., 1992] B. Selman, H. Levesque, and D. Mitchell. A new method for solving hard satisfiability problems. In *Proc. AAAI-92*, 1992.

[Smith, 2003] D. Smith. Special issue on the 3rd international planning competition. *JAIR*, 20, 2003.

[Stanovich, 2004] K. Stanovich. *The Robot's Rebellion: Finding Meaning in the Age of Darwin*. Chicago, 2004.

[Sutton and Barto, 1998] R. Sutton and A. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1998.

[Tooby and Cosmides, 1992] J. Tooby and L. Cosmides. The psychological foundations of culture. In J. Barkow, L. Cosmides, and J. Tooby, editors, *The Adapted Mind*. Oxford, 1992.

[Vidal and Geffner, 2005] V. Vidal and H. Geffner. Solving simple planning problems with more inference and no search. 2005.

[Zajonc, 2004] R. Zajonc. *The Selected Works of R.B. Zajonc*. Wiley, 2004.

Building Plans for Household Tasks from Distributed Knowledge

Chirag Shah* and **Rakesh Gupta**
Honda Research Institute USA, Inc.
800 California Street, Suite 300
Mountain View, CA 94041
chirag@cs.umass.edu, rgupta@hra.com

Abstract

To accomplish a household task, an autonomous system needs a plan with steps. It is desirable to derive this plan dynamically instead of pre-coding it in the system. In this paper, we find a plan by using common sense knowledge collected from volunteers over the web through distributed knowledge capture techniques. This knowledge consists of steps for executing common household tasks.

We first pre-process the data with part-of-speech (POS) tagging to identify the actions and objects in the steps in all available plans for the task. We then determine the order of the steps to accomplish the task using discriminative as well as generative models. For the discriminative approach, we cluster the plans using hierarchical agglomerative clustering and choose a plan from the biggest cluster. In the contrasting approach, we make use of generative Markov chain techniques. Using human judgments, we show that the generative model with the first order Markov chain has the best performance. We also show that environmental constraints can be incorporated in the generated plans.

1 Introduction

The long term goal of our work is to make indoor mobile robots that work in environments like homes and offices more intelligent through common sense. Mobile robots in homes and offices will be expected to perform tasks within their environment to satisfy the perceived desires and requests of their users. In order to meet these needs, we want to endow these robots with some knowledge to use as the basis to accomplish these tasks. Examples of household tasks include making coffee, washing clothes, and cleaning a spill.

Common sense does not require expert knowledge, and hence it may be gathered from non-specialist net citizens (netizens) in the same fashion as the projects associated with the rather successful *OpenMind Initiative* pioneered by David Stork [Stork, 1999]. In this paper, we focus on how we can use the knowledge collected about task steps in the OpenMind Indoor Common Sense project (OMICS) [Gupta and Kochenderfer, 2004] to populate an initial robot knowledge base with default household task plans.

In the work reported here, we show two approaches to find a plan for a given task. Our first approach is *discriminative*, where we perform hierarchical agglomerative clustering of the plans before choosing one. Our second approach is *generative*, where we make use of the first order Markov chains. We start with all the plans from OMICS database for the selected tasks and capture the sequence information between steps using the first order Markov chains. This allows us to combine multiple plans with interleaved chains into a generative model from which an appropriate plan can be derived. We also experimented with capturing the globally optimal sequence of steps.

Figure 1 shows an outline of different techniques. Technique two is discriminatory, and techniques three, four and five are generative. All these techniques are compared to the baseline technique of selecting a random plan (technique one). We performed experiments with human subjects and used statistical significance tests to compare these techniques.

The rest of the paper is organized as follows. The following section discusses the related work. We then give examples of data in our knowledge base and describe preprocessing to extract action-object pairs. Section 4 discusses discriminative approach along with the clustering technique to find the largest cluster of plans reflecting consensus. We then describe the generative approach, for capturing the sequence information using the first order Markov chains to generate locally and globally optimal plans. We also show how the generative approach can be used to generate a plan with environmental constraints. The results and analysis are presented in section 6. We finally conclude our paper with some pointers to the future work.

2 Related Work

In the past, expert systems have been used to encode the steps for accomplishing a task algorithmically [Waterman, 1986]. A key component was the capture of human expert knowledge using a laborious manual process. However, not everything that humans learn is taught by the experts. Most of the activities of our day-to-day life are learned by observations and experience – by looking at other non-experts (e.g. tying shoe laces).

*Currently at University of Massachusetts, Amherst, MA

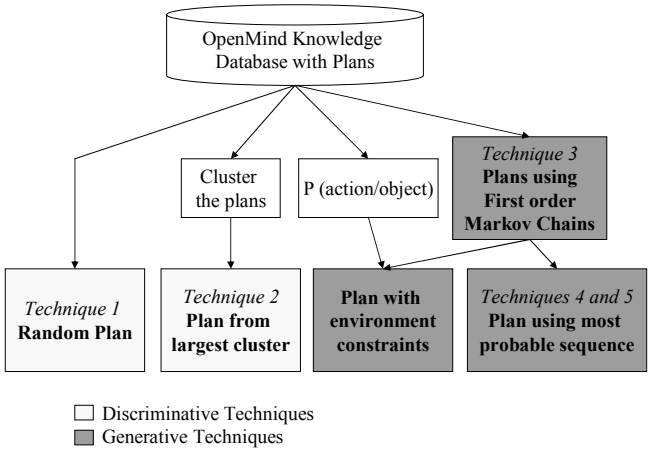


Figure 1: outline of proposed approaches.

Human actions can be analyzed on a variety of different levels. At lower level, execution of an action can be controlled by a motor response schema of sensory motor mapping [Schmidt, 1975]. For higher levels of action control, concepts such as scripts [Schank and Abelson, 1977] and memory Organization Packets (MOP) [Schank, 1982] have been proposed to represent the organization of well-learned activities such as going to a restaurant or visiting a doctor for surgery. When a MOP is activated, only one of the steps is generally carried out at a time, but steps can sometimes be combined with other activities (for example, one can read while waiting at doctor's office).

Between these low and high level extremes lies a whole range of well-learned activities like making breakfast, cleaning one's teeth, starting a car, dressing and so on. In these situations, the steps are represented as discrete units and their parallel execution (although possible) rarely occurs. Thus in toothbrushing, one can delay or do something between putting toothpaste on the toothbrush and brushing one's teeth, it is not generally advisable.

At this mid-level, Cooper and Shallice [2000] presented a computational model for selection of of steps for routine tasks based on competitive activation within a hierarchically organized network of action schemas. Their activation model for sequential step selection was based on the Contention Scheduling theory of Norman and Shallice [1980]. This model was demonstrated in the routine task of preparing coffee. Under normal functioning, the model was able to generate a sequence of simple actions (pick up spoon, dip spoon in sugar bowl etc.) culminating in a drinkable cup of coffee.

According to Rasmussen *et al* [1983], human activity in such routine tasks is oriented towards a goal and controlled by a set of rules which have proven successful in the past. The sequence of steps is typically derived empirically during previous occasions, communicated from another person's knowhow or a cookbook recipe. We are interested in capturing such procedures for household tasks.

In contrast, literature and work in AI planning [Weld, 1999] falls under goal controlled exploratory behavior category. Here attempts are made to reach the goal using knowledge of different plans and a successful sequence is selected. Such planning is complementary to our work and is critical during execution of these tasks.

One source of common sense knowledge is the web. For instance, web-sites such as *eHow*[1] list the steps to perform activities[2]. Intel developed a system called *Probabilistic Activity Toolkit (PROACT)* to build activity models [Philipose *et al.*, 2003]. They automatically identified activities by observing the objects involved in the activity [Perkowitz *et al.*, 2004]. They found the relevance of various terms to a given activity from the web. For instance, the word *cup* is highly related to activity *making tea* because *cup* occurs frequently on the web-pages about *making tea*.

This idea of using the web as the information source is very attractive. However, while extracting knowledge from the web, one has to deal with high variance and noise in web information and large documents. We have better information on steps for household tasks from Open Mind Indoor Common Sense (OMICS) database. In OMICS, volunteers are prompted with household tasks and asked for steps to accomplish such a task. We still have to extract semantic information from the steps, as well as deal with issues of noise and consistency in the data.

3 Semantic Data Extraction

There are more than 150 tasks in OpenMind database like *making coffee, cleaning the floor, washing clothes* for which plans have been entered by the users. Each task in our database consists of a number of plans and each plan consists of a sequence of instructions. A sample plan from the database is shown below:

```
Task: wash clothes
Steps: collect clothes
       move to washing machine
       place clothes in washing machine
       add detergent to clothes
       close lid of washing machine
       start washing machine
```

Our objective is to make use of these plans for the given task to build a model. From this model. we can extract a plan, or derive a plan (which may not be present in the original set of plans), or generate a custom plan based on environmental constraints. In this section, we further describe how this data is pre-processed by extraction of action-object pairs.

3.1 Extracting action-object pairs

We extract the action and object from a given step for better processing downstream. For extracting action-object pairs we first parse the instruction with Brill's part-of-speech (POS) tagger [Brill, 1992]. We then identify the first verb as the action. If the verb is followed by a proposition, we combine the preposition with the action. Finally, we identify the first noun phrase as the object of the action. Since most of the steps in the tasks are instructional in nature, this simple procedure performs surprisingly well. The result of parsing the above plan is shown below:

[1] http://www.ehow.com

[2] These activities correspond to our tasks.

```
Task: wash clothes
Action object pairs for steps:
        collect, clothes
        move to, washing machine
        place, clothes
        add, detergent
        close, lid
        start, washing machine
```

We are also interested in finding how the objects are related to various actions for every task. We, therefore, find the following conditional probability distribution for every task.

$$P(action|object) = \frac{f(action, object)}{f(object)} \quad (1)$$

where $f(action, object)$ is the number of times $action$ occurs with $object$ and $f(object)$ is the number of times $object$ occurs in the given task across all the plans.

4 Discriminative Approach

A simple way of coming up with a plan is to select a random plan from the set of the plans in the database. This forms our baseline. In the discriminative approach, we select a plan from a subset representing the majority consensus. Majority consensus is reflected by the biggest cluster of plans in the task.

We perform hierarchical agglomerative clustering where we group similar plans, and merge similar groups into larger groups [Salton, 1989]. For each task, we find the similarity between two plans p_i and p_j as the following.

$$Sim(p_i, p_j) = \frac{len(largest\ matching\ sequence)}{len(p_i)len(p_j)} \quad (2)$$

where $len(p_i)$ is the individual number of steps in the plan p_i and $len(largest\ matching\ sequence)$ is the number of common steps in plans p_i and p_j. When individual plans are compared we use the similarity criteria as given in the equation above, but when clusters are compared, we use group average to compare clusters on the basis of average similarity. In our system, we can specify the desired number of clusters. We empirically found that plans for our tasks fall in five or fewer categories. Therefore, we choose to have at the most five clusters of plans for each task.

We have found these clusters to correspond to distinct techniques for accomplishing the task. For example for *making coffee* task, the different clusters correspond to using coffee maker, instant coffee, and expresso machine. Since a majority of people entered plans to make coffee using the coffeemaker, that cluster was the largest. After clustering, we randomly select a plan from the largest cluster.

5 Generative Approach

So far we have selected a plan from the original set of plans. In this section, we describe how we can instead *generate* a plan. We propose to construct a model for each task, called *Task Model*, using the first order Markov chains [Rabiner, 1989]. The primary motivation for the Markov chain is the inherent sequential nature of steps in a given plan. We model each plan as a first order Markov chain, where each step depends on the previous one with no hidden states.

5.1 Constructing the Task Model

All the plans are encapsulated between the start and end steps. To build the Task Model, we link the first step to the start state with probability 1.0. We keep linking all other steps in order with probability 1.0, ending in the end state. We repeat this for all plans in the database with appropriate probability computations. For example, we initially have a transition from A to B with probability 1.0. When we get a new transition from A to C, we create an additional link from A and recompute the probabilities of links from A as 0.5 each. It is possible for the same instruction to occur more than once in the model. For instance, in case of *wash the clothes* task, *open the lid* can occur before putting the clothes in the washing machine and after the washing is done to take out the clothes.

After executing the above procedure for *washing clothes* task we combine all these links to generate a graph. All these states are represented in the graph as nodes and joined according to their transition probabilities. A sample construction is shown in figure 2.

There are various advantages of building such Task Model. Firstly, we do not have to store all individual plans for a task. We store a model for each task which makes the storage space linear in the number of steps, rather than linear in the number of total plans in the database. Secondly, models evolve with technological advancements and encompass new information as newer plans for tasks come up. When we have a new version of the database of plans entered by non experts, we can either generate a new model from the data using the same procedure described here or update the existing model with new probabilities and new states. Finally, having a model allows us to generate consensus plans and more complex plans that do not exist in the database or use available objects in the environment.

The following subsections discuss details of how the generated Task Model can be used to derive a plan with the different generative techniques. We first describe technique three, followed by techniques four and five using the most probable sequence.

5.2 Plans using the first order Markov Chains

We have already captured step sequence information in our Task Model. To derive a locally optimal plan, we go through the most probable sequence of steps. At every state starting from start, we choose the next state as the one with the highest probability. The state at time t is found using the following equation:

$$NextState(t) = \arg\max_{s_i} p(s_i|s_j) \quad (3)$$

where t is the current time step, s_j is the state at time $t-1$, and s_i are all the successor states of s_j. To avoid cycles, we remove all the incoming links to the step that we visit[3]. A

[3]Note that this will not prevent us from using the same instruction again in the plan generation. This is because the same instruction may be represented by more than one states. For instance, "close the lid" instruction may be represented by two states - one that occurs before "put the clothes" state and the other that occurs after "take out the clothes" state.

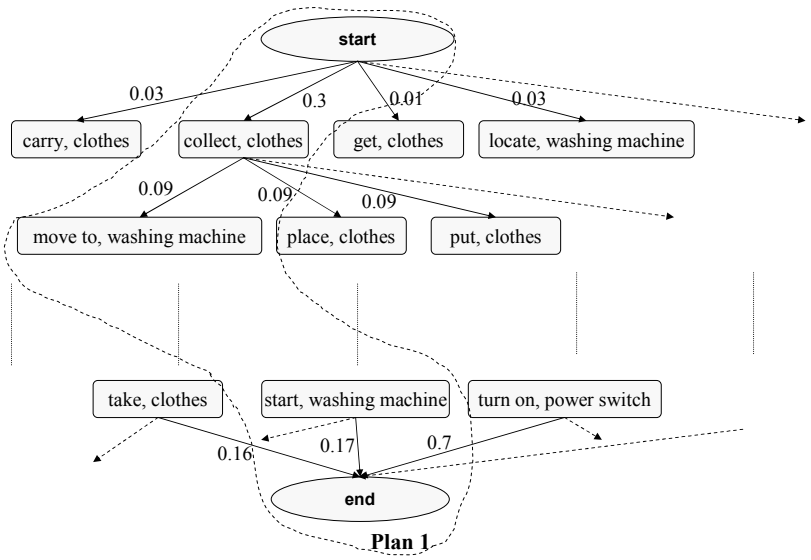


Figure 2: Portion of a Task Model. The dashed outline shows one of the plans.

sample output for the task *wash the clothes* is given below. It is in the format *step* → *transition probability* → *step*.

```
start -> 0.36 -> get, clothes
get, clothes -> 0.3333 -> locate, machine
locate, machine -> 0.4285 -> move to, machine
move to, machine  -> 0.2727 -> fetch, clothes
fetch, clothes -> 0.5 -> open, machine
open, machine -> 0.8 -> put, clothes
put, clothes -> 0.5294 -> add, detergent
add, detergent -> 0.5 -> close, machine
close, machine -> 0.2857 -> start, machine
start, machine -> 0.6666 -> end
```

Putting these steps together, we get a plan. The evaluation of its *goodness* is discussed in the next section.

5.3 Generating globally optimal plans

So far we have considered dependency of a state on the previous state to reduce the computation. For global optimality, we consider the relation of present state to all the previous states. Therefore, our formulation of selecting the next state as given in equation (3) changes to the following:

$$NextState(t) = \arg\max_i p(s_i|s_1, s_2, ..., s_{i-1}) \quad (4)$$

where t is the time step on which we are trying to find the best state, $s_1, s2, ..., s_{i-1}$ are the states that occurred till $t-1$ and are linked through a path in the model, and i iterates over all the possible states at step t.

To implement this idea, we computed the probability along different possible sequences from start to end in our Task Model. The sequence that gives the maximum probability is chosen for the plan. In technique 4, we select a random sequence if there are multiple choices with the largest probability. In technique 5, we select the shortest of these sequences if there is a tie. The results of all these techniques along with the analysis are given in the Results section.

5.4 Plan with environmental constraints

In a real-time system, we want to consider the environmental constraints while generating a plan. Such constraints may be provided by the user or obtained by the system using sensors. Sensors may provide information about what objects are available in the environment. However, our data is represented in the units of action-object pairs. To convert the restrictions on objects to action-object pairs, we assume that the most likely action is one that occurs most frequently with the object. We make use of the $P(action|object)$ probabilities that we found earlier in equation (1). For handling a constraint to use a given object in the task, we find the most probable action to be associated with that object:

$$\arg\max_{action} P(action|object) \quad (5)$$

In plan generation, we choose children of the current step with restrictions. If there is more than one choice, we choose the one with highest probability. If there are no constraints for a step, we choose the one with the maximum probability on its link. In order to avoid cycles, we remove all the incoming links at each visited step.

It is important to note here that even though we are associating the observed objects to their most likely actions, these actions are not forced in the plan. It is possible that the plan generation process neglects such steps to be consistent with the rest of the steps. A sample output for the *washing clothes* task with restriction to make use of *water, clothes,* and *washing machine* is given below:

```
Found restriction: "feed, water" with
    probability 0.2
Found restriction: "put, clothes" with
    probability 0.32
Found restriction: "start, washing machine"
    with probability 0.15
The plan:
  start -> 0.36 -> get, clothes
```

```
get, clothes  -> 1 -> put, clothes
put, clothes  -> 1 -> feed, water
feed, water  -> 1 -> feed, detergent
feed, detergent  -> 1 -> set, timings
set, timings  -> 1 -> start, washing machine
start, washing machine  -> 0.6666 -> end
```

This plan is different from the one that we derived earlier without any constraints. Because of this method's bias towards choosing the step that fulfills one of the restrictions, the corresponding transition probabilities are set to 1.0.

6 Results and Analysis

We selected 105 tasks, each with at least 25 plans in the OMICS database and compared the following techniques:

1. Baseline: a plan selected randomly.
2. Discriminative approach: a plan selected from the largest cluster.
3. Generative approach: a plan generated from the corresponding Task Model.
4. A plan generated from the Task Model by considering the probability of the whole sequence.
5. A plan generated from the Task Model by considering the probability of the whole sequence, and choosing a plan with minimal length if there is a tie.

All these techniques made use of the same data and preprocessing. In order to compare these techniques[4], we based our evaluation on the following criteria:

1. *Completeness.* Plan for a task should be complete. A plan for cleaning the floor, which spilled the water and the soap on the floor but did not mop it to dry, would be an incomplete plan.
2. *Correct sequence.* The sequence of steps should be consistent. A plan, which poured coffee from carafe into a mug before adding water to coffee-maker, would be rated low using this criteria.
3. *It should make sense.* For example if a coffee making plan used both a coffee-maker and instant coffee, it would be rated low.
4. *Not too many details of interaction with objects.* Given two plans with same number of objects, a plan with higher level description is preferable to low level detailed instructions. For making coffee for the step of adding filter to coffee-maker,
   Preferable: add filter
   Less preferable: find filter, take one filter, add filter

We asked 10 users to rank five plans for each of 105 household tasks. This evaluation took users about 2 hours to fill. For every task we averaged over the 10 evaluations for each of the 5 plans. The maximum score for a technique for a given task would be 5 (the worst) and the minimum score would be 1 (the best). We then add these scores for all the 105 tasks for each technique. Table 1 summarises the results, from which we make the following observations:

Technique	Name	Avg. score	Overall ranking
1	Baseline	267.63	5
2	Discriminative	257.45	3
3	First order Markov chain	245.90	1
4	First order Markov chain with full sequence	250.09	2
5	First order Markov chain with full shortest sequence	266.00	4

Table 1: Results of the user judgments

- Selecting a plan randomly from a given set of plans (baseline) gives the worst performance. This indicates that if we rely on the *knowledge* of just one person or one source, then we have high likelihood of a bad plan.
- Technique 2, which is a discriminative approach, does better than the baseline. A random plan from the consensus cluster is better than a random plan.
- Technique 3 using first order Markov Chains does the best. This is a generative model that *learns* the *knowledge* from the given data and generates a plan. This approach is attractive for a number of reasons. First, it considers a step as a unit instead of a plan, thus not confining itself to a particular plan like the first two techniques. Second, it is able to remove some noise and spurious data through the process of learning. Third, it captures the consensus at the level of steps and their sequence.
- Techniques 4 and 5 do not do as well as 3. We believe the main reason is the lack of sufficient number of plans required to perform the inferencing on long sequences of steps.

The scores reported in table 1 gives an idea of the performance of various techniques. To compare these techniques, we performed a paired two-tailed t-test. This test determines if the outcome of two different techniques come from the same distribution. If the original distributions are significantly different, then the one that provides better results is said to be performing significantly better over the other. This statement about significance is associated with a level of confidence. The p-values for different techniques are given in table 2. We can see that with the confidence interval of 95%, techniques 3 (locally optimal plan) and 4 (globally optimal plan) perform statistically significant over the baseline. Both of these techniques are based on the first order Markov chain generative models.

7 Conclusion and Future Work

In this paper we discussed the problem of selecting or extracting a plan to perform a task from the given set of plans that have been collected by distributed knowledge capture techniques. We proposed a discriminative as well as generative

[4]The implementation that made use of the environmental constraints was difficult to judge and evaluate; therefore, we did not include it in our evaluation. However, it is a special case of technique 3 and inherits advantages of that technique.

Technique	p-value
2	0.0838
3	0.0004
4	0.0036
5	0.7866

Table 2: p-values from paired two-tailed t-test. The tests were done by making pairs of ranks given by baseline and the technique specified in the first column.

approaches to derive a plan. We used hierarchical agglomerative clustering for the former approach and first order Markov Chains for the latter one.

Although the discriminative approach gave better results than the baseline, the generative approach did even better. For the generative approach we first construct a Task Model from the available plans. We use First Order Markov Chains to find the most likely sequence of steps. We can also incorporate information about available objects in the plan generation process. Our experiments showed reasonable plans as outputs in discriminative as well as generative models. Since the goodness of these results cannot be measured objectively, we used human subjects to evaluate our results. We also showed that the differences among techniques were statistically significant.

In our plan representation and generation processes, we perform shallow Natural Language Processing (NLP) for finding action-object pairs. Further NLP can improve the results. For instance, we can find the synonyms of the actions as well as objects and merge them to simplify the Task Model before generating the plans. In future work, we also plan to automate the specification of environmental constraints for objects in the environment using RFID or vision sensors.

8 Acknowledgments

This work was done while Chirag Shah was a summer intern at Honda Research Institute USA, Inc. Thanks are also due to anonymous reviewers and the users of the Open Mind Indoor Common Sense web-site for their data and feedback.

References

[Brill, 1992] Eric Brill. A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento, IT, 1992.

[Cooper and Shallice, 2000] Richard Cooper and Tim Shallice. Contention scheduling and the control of poutine activities. *Cognitive NeuroPsychology*, 17(4):297–338, 2000.

[Gupta and Kochenderfer, 2004] Rakesh Gupta and Mykel Kochenderfer. Common sense data acquisition for indoor mobile robots. In *Nineteenth National Conference on Artificial Intelligence (AAAI-04)*, July 25-29 2004.

[Norman and Shallice, 1980] D. Norman and T. Shallice. *Attention to Action: Willed and automatic control of behavior*, pages 1–18. Plenum Press, New York, 1980.

[Perkowitz *et al.*, 2004] Mike Perkowitz, Matthai Philipose, Kenneth Fishkin, and Donald J. Patterson. Mining models of human activities from the web. In *Proceedings of the 13th Conference on World Wide Web*, pages 573–582. ACM Press, 2004.

[Philipose *et al.*, 2003] Matthai Philipose, Kenneth P. Fishkin, Mike Perkowitz, Donald Patterson, and Dirk Haehnel. The probabilistic activity toolkit: Towards enabling activity-aware computer interfaces. Technical Report IRS-TR-03-013, Intel Research Laboratories, November 2003.

[Rabiner, 1989] Lawrence R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, February 1989.

[Rasmussen, 1983] Jens Rasmussen. Skills, rules and knowledge: Signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-13(3):257–266, May/June 1983.

[Salton, 1989] Gelad Salton, editor. *Automatic Text Processing: The transformation, analysis, and retrieval of information by computer*. Addison Wesley, 1989.

[Schank and Abelson, 1977] R. C. Schank and R. Abelson. *Scripts, Plans, Goals and Understanding*. Lawrance Erlbaum Assoiates Ltd., Hove, UK, 1977.

[Schank, 1982] R. C. Schank. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, Cambridge, 1982.

[Schmidt, 1975] R. A. Schmidt. A schema theory of discrete motor skill learning. *Psychological Review*, 82(4):225–260, 1975.

[Stork, 1999] David G. Stork. The Open Mind Initiative. *IEEE Expert Systems and their Applications*, 14(3):19–20, May/June 1999.

[Waterman, 1986] D. A. Waterman. *A Guide to Expert Systems*. Addison Weseley, 1986.

[Weld, 1999] Daniel S. Weld. Recent advance in ai planning. *AI Magazine*, 20(2):93–123, Summer 1999.

Innate Planning Mechanisms

Sule Yildirim
Complex Adaptive Organically-Inspired Systems Group
Department of Computer Science
The Norwegian University of Science and Technology, Trondheim, Norway
yildirim@idi.ntnu.no

Abstract

In this paper, we will discuss whether there could be any means to bridge the gap between the Symbolic and Subsymbolic AI. One way to do this is to ask ourselves if human brain executes any planning algorithms. We see that we have taken a series of steps when we are done with planning in a situation. Taking a series of steps during planning might be a result of the execution of an innate planning algorithm. If we really are executing a planning algorithm, then we believe, its function is very general and is to set the conditions which will trigger a next step to take. A step to take might be the execution of an IF rule as an example. IF rule executions are not the only steps to take while planning, however, for simplicity, they are assumed as the only ones here. There is not any neuro-scientific evidence against the possibility that human mind incorporates an innate planning algorithm that triggers the next rule to execute (the step to take) yet. For that reason, in this paper we will investigate that possibility.

Keywords: Action sequencing, action associations, concept-action associations, emergence of association mechanisms

1 Introduction

Classical AI symbol systems are criticized basically for two points [Steels, 1996]:

1) Their problem solving functionality such as planning needs to be programmed by hand as opposed to evolving adaptive intelligent systems.
2) The symbolic descriptions of the reality need to be given to them.

As a result, there are already many studies where adaptive intelligent systems are evolved as opposed to being hand designed [Nolfi and Floreano, 2000]. There are also studies which reject the presence of any kind of representations [Brooks, 1991].

In [Steels, 1996] it is stated that most of the work assumes that there are abstraction facilities in neural networks or a new higher level dynamics that may emerge. However, none of the systems developed in this studies are yet able to achieve the high level capabilities of human beings such as planning and reasoning.

In this paper, we will consider the possibility of having an innate planning algorithm that sets the conditions which will trigger a next step to take. We will mainly consider task planning but not motion planning and navigation. In robotic literature, task planning is defined as the planning activity that calculates the order in which a robot should execute "actions" or "sub-tasks", in order to reach a specified goal. Assembly and "travelling salesman"-like jobs go into the task planning category.

The idea of having the innate algorithm is similar to the idea of having a traversal algorithm in Symbolic AI because a traversal algorithm, although is not as much general in function as the innate planning algorithm we are thinking of, shows a way to trigger the next action or step to take also.

An innate task planning algorithm might be what we need to borrow from Symbolic AI and if we do so then we can direct our studies to emerge the innate planning algorithm.

In Section 2, we will elaborate on the presence of rules in human mind with a movie planning example. Section 3 will talk about association of concepts with each other and with the rule in execution. We will explain the composition of the innate planning algorithm in this section also. Section 4 will be our conclusions.

2 Presence of Rules in Mind

The sentences we encounter either on paper, on computer screens or in spoken language are analyzed syntactically before we can actually get meanings out of them. This is managed by us using a set of grammar rules which have their mental representations [Jackendoff, 1993; Pinker, 1993].

We many times per day experience ourselves applying grammar rules while forming sentences. This becomes more obvious when we learn a new language. Although [Jackendoff, 1993] suggests a universal grammar inherited genetically in addition to other steps of learning a language, our point here is to keep attention to the fact that if we have

grammar rules in our minds then we might have rules other than grammar rules represented in our minds as well.

How could it be inferred that we have other rules than grammar rules represented in our minds and what other importance does this have in addition to knowing we represent rules as well as symbols in our minds?

If we knew we did execute rules in our minds, we would be closer to inferring that we might also be capable of executing planning algorithms in our minds.

When I think about rather seeing a movie than attending a party this evening, the questions that occupy my mind are possibly the kind of movies that are on show tonight, whether there being any science fictions movies on or not, whether I have already seen them or not, whether the show times are too late or not and some more whether questions if not What, Which, Where and How questions. A "whether" follows a previous "whether". For practical purposes we will replace a "whether" with an "If" from this point on. It is true that one "if" might remind me of another "if". As an example, I might ask if there are any science fiction movies on and then this might remind me of a science fiction movie I have already seen and of its director and I might start wondering if I could find a movie for tonight which is directed by the particular director. However, regardless of one if leads me to another if or to other thoughts, it is definite that I am applying an IF rule as a part of my thinking process. These rules are represented in my brain in the form of various biological neuron patterns.

Planning for tonight is mostly dependent on our beliefs (what we already know about our world although our interpretations of the world might be different from reality), desires and intentions which are internal at the time of thinking although they might be produced as a result of earlier interactions of human beings with the real world. For example having already seen a good movie of John, we can believe that John is a good director. This particular belief necessitates having representation of the real person John internally as well as having the internal representations of concepts "good" and "director". Thinking that "John is a good director" demands us to refer to the concepts of "director" and "good" explicitly as well as the real person John at the time of thinking.

If we now go back to our discussion of having If rules represented in our minds, we can say that it is the case that first we find answers to conditional parts of an "IF" expression, such that we can apply the THEN part of the expression (rule). However, it is also possible that because of lack of information, we might suspend working on the conditional part of an IF rule and jump to another if rule or another thought. The execution of the new if rule might supply us with enough information to resolve the previous if rule so we can go back to the execution of the previous one and complete it. If we continue like this, we might come up with a list of things or ideas that we would want to achieve.

In the movie example, I might finally decide to go to the party instead, if I conclude that none of the movies on the show are interesting. On the other hand, I might have decided to go to see one of the movies at 7 o'clock but after having dinner in a nearby restaurant and yet meeting with a colleague in the mathematics department to deliver him his book I borrowed a week ago before that.

As a result, human brain actually might be acting like a computer which executes the steps of an algorithm while executing IF rules.

The statement that brain acts like executing the steps of an algorithm is a metaphor to the execution of an IF rule, suspending the execution of an IF rule and jumping to another IF rule, going back to the execution of the previous IF rule and so on. Each rule in the execution sequence can be inspired by the other and hence appear and take its term in the whole thought process.

If we consider all of the rules that are invoked during planning as a part of a rule search space, then there can be an *algorithm,* which decides which rule is followed by which rule.

It does not seem to be a mistake to consider each of these rules as corresponding to a node of a search tree in symbolic AI. We can also use symbolic AI tree search strategies (i.e. depth first, breadth first) as a metaphor for the type of algorithm we mention here. The algorithm can trigger other rules for execution than the one which is now in execution. It is possible that the first rule is triggered by a problem from the environment as well as by internal beliefs, desired or intentions.

Although planning, decision making or thinking happen in the frontal lobe, they are in tight communication with other parts of the brain in terms of retrieving other rules or symbols (from memory), sending back newly inferred rules and symbols (to memory), making associations between rules and concepts, activating motor cortex and other possible handling (actions).

3 Rule and Concept Associations

In a situation of making a decision, as above, between attending a movie and a party, IF rules seem to be applied and one IF rule seems to lead to another IF rule.

Following statement can be the very basic algorithm of our minds which invokes the next rule to execute:

"Execute the next associable rule while resolving the current rule or after the execution of the current rule is finished and do concept associations meanwhile".

This algorithm resembles a one step traversal algorithm that can be applied on a search tree but it is more general and since it considers associations of the current rule to other rules and concepts, it is situated in the sense that these rules and concepts are exposed to updates from environment.

In order to achieve the statement of the algorithm, we could possibly have yet other rules which actually form the algorithm itself. We will call these rules as meta rules to

separate them from other IF rules. An example "meta IF rule" could tell our mind how to execute an "IF rule" as follows:

"Execute the preconditions of an IF rule first and then execute the THEN part".

If several IF rules are invoked by real world problems at the same time for simultaneous thoughts, the meta rule can be applied to each IF rule and two or more rule executions can take place in parallel.

Given the message of executing the preconditions first, the possible associations with the current rule and the other rules and concepts will be achieved.

From dynamical systems perspective, the meta rules will correspond to the laws of change [Holland, 1998] because these rules create the dynamism to execute new rules and make associations.

Concepts that are associated with IF rules refer to the mental states as described in [Dorffner, 1992] in our work. Concepts have non-linguistic representations and they are shaped by context and experiences of an agent who is forming those concepts. We extend this description of concepts to IF rules and assume them to be mental states also.

In Figure 1, an example for concept-concept associations is given. In the figure, "Movie" is a concept which has features such as "Name", "Date", "Time", "Seen before", "Type" and "Director". These features are also concepts. Each feature (concept) can have possible different values. For example date can be "Thursday" or "Tuesday". In fact, we see each feature value as a concept also.

The links between feature concepts to value concepts are absent or present depending on what value another feature has. For example, if "Name" feature has the value of "ET", "Date" feature will have a value of "Tuesday". However, if it is "XY", "Date" feature will have a value of "Thursday".

In this example, features, values and the "Movie" concept are part of a Movie context. However, "Date" and "Time" features can be part of another context such as delivering a book to the mathematics department on Thursday and before 4 pm. I might switch to the context of delivering a book while I am thinking about the date of a movie because "Date" feature is also part of the book context (Figure 2).

I can switch back to the Movie context when I decide that I should deliver the book today because it is Thursday. One of the movies, as an example, XY will supply all my conditions of seeing a movie tonight because it will activate the feature concepts of "Science fiction", "Thursday", "No", "John" and "7 pm". All these features can be connected by a node which represents an "And" concept and can lead to another concept which is "See the Movie". That is, we actually are executing and if rule which is:

IF (Type = "Science fiction" and Date = "Thursday" and Seen before = "No" and Director = "John" and Time = "7 pm") THEN See the Movie.

In the figures, arrows do not represent the spread of activations. They only represent which concept is related to which concept. Feature values given in the figure are the possible values only for the Movie context.

On the other hand, while building symbolic planning systems, researchers have encountered many problems such as frame problem, temporal projection problem etc. We ignore those problems and how their solutions could be within our work because we are not aiming at building a planning system that can plan as well as or better than the existing symbolic planners but we are questioning whether there can be any innate planning algorithms or not and if so what their role could be in human mind.

Finally, we will suggest that the innate planning algorithm is nothing but the application of the Hebb's rule [Hebb, 1949]. That is, some of the concepts in our minds are activated because of either external events or internal beliefs, desired and intentions. On the other hand, that activated concept or concepts activate another one depending on how strong or weak a link (synapse) between the current concept and the next concept is. Concepts can be part of rules and thus activation of another concept might mean activation of another rule.

4 Conclusions

We believe that existing research in artificial neural networks [Kohonen, 1984; Kosko, 1988], evolutionary computation [Nolfi and Floreano, 2000] and others [Prescott *et al*., 2002] can be scaled up to form a computational model of human mind where the components of the model are rule representations, concept representations, concept-concept associations, rule-rule associations and belief, desire and intention representations. There are already studies in this direction [Cangelosi, 2004].

We also aimed at pointing to a similarity in terms of task planning between Symbolic AI task planning systems and a neural network task planning system which can be like the one presented in this paper. We believe that this similarity which points to a navigation in a rule search space while planning a task could be one of the means of bridging the gap between Symbolic and Subsymbolic AI.

The next step for us will be the implement the system described in this paper.

Acknowledgments

We thank to Prof. Dr. Keith Downing and Dr. Diego Federici from CHAOS group who read the earlier drafts of this paper and commented on them in detail. We also thank to the anonymous reviewers from whose reviews we benefited greatly.

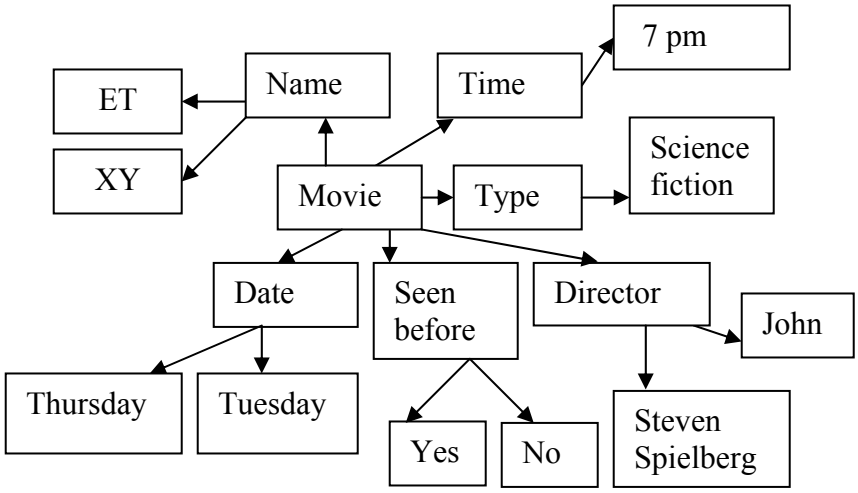


Figure 1. Concept to concepts associations

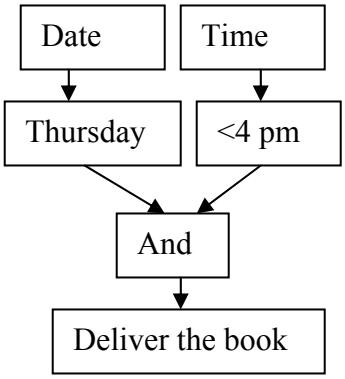


Figure 2. Concepts "Date" and "Time" remind of delivering the book.

References

[Brooks, 1991]Rodney A. Brooks. Intelligence Without Representation. *Artificial Intelligence Journal* (47), pages 139–159, 1991.

[Cangelosi, 2004] Angelo Cangelosi. The sensorimotor bases of linguistic structure: Experiments with grounded adaptive agents. In S. Schaal et al. (eds.), *Proceedings of the Eighth International Conference on the Simulation of Adaptive Behaviour: From Animals to Animats 8*, Cambridge MA, MIT Press, pages 487-496, 2004.

[Dorffner, 1992] Georg Dorffner. A Step Toward Sub-Symbolic Language Models without Linguistic Representations. In Reilly R. and Sharkey N. (eds.), *Connectionist Approaches to Natural Language Processing*, Vol. 1, Lawrence Erlbaum, New Haven/Hillsdale/Hove, 1992.

[Hebb, 1949]. Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. New York: Wiley, 1949.

[Holland, 1998] John H. Holland. *Emergence From Chaos To Order*. Helix Books, 1998.

[Jackendoff, 1993] Ray Jackendoff. *Patterns in the Mind: Language and Human Nature.* Harvester Wheatsheaf, 1993.

[Holland, 1998] John H. Holland. *Emergence From Chaos To Order*. Helix Books, 1998.

[Kohonen, 1984] Teuvo Kohonen. *Self Organization and Associative Memory.* Springer Verlag, Berlin, 1984.

[Kosko, 1988] Bart Kosko. Bidirectional Associative Memories. *IEEE Transactions on Systems, Man, Cybernetics SMC-L8*, pages 49-60, 1988.

[Levine, 1998] Daniel S. Levine. *Explorations in Common Sense and Common Nonsense*. On-line book. 1998.

[Nolfi and Floreano, 2000] Stefano Nolfi and Dario Floreano. *Evolutionary Robotics*. MIT Press, 2000.

[Pinker, 1993] Steven Pinker. *How the Mind Works*. W. W. Norton & Company, Inc., 1993.

[Prescott *et al*., 2002] Tony J. Prescott, Kevin Gurney, Fernando Montes-Gonzales, Mark Humphries, and Peter Redgrave. The Robot Basal Ganglia: Action Selection by an embedded model of the basal ganglia. *Basal Ganglia* 7, Plenum Press, 2002.

[Steels, 1996] Luc Steels. The origins of intelligence. In *Proceedings of the Carlo Erba Foundation Conference on Artificial Life*, Milano, Fondazione Carlo Erba., 1996.

Ecological Integration of Affordances and Drives for Behaviour Selection

Ignasi Cos-Aguilera[†], **Lola Cañamero**[‡], **Gillian M. Hayes**[†] and **Andrew Gillies**[†]
[†]School of Informatics, University of Edinburgh,
5 Forrest Hill, Edinburgh EH1 2QL, Scotland, UK.
[‡]School of Computer Science, University of Hertfordshire
College Lane, Hatfield, Herts, AL10 9AB, UK.

Abstract

This paper shows a study of the integration of physiology and perception in a biologically inspired robotic architecture that learns behavioural patterns by interaction with the environment. This implements a hierarchical view of learning and behaviour selection which bases adaptation on a relationship between reinforcement and the agent's inner motivations. This view ingrains together the basic principles necessary to explain the underlying processes of learning behavioural patterns and the way these change via interaction with the environment. These principles have been experimentally tested and the results are presented and discussed throughout the paper.

1 Introduction

The problem of behaviour selection, ergo knowing *what to do next* has been approached by different disciplines, from neuroscience to robotics [Dayan and Balleine, 2002; Ávila García and Cañamero, 2002].

Animal perception responds to the principle of ecology which directly conditions the way they learn and select behaviours. An analogous perception has been modelled and integrated in a framework to *learn behavioural patterns* inspired on a double hypothesis of learning [Schultz et al., 1993] and behaviour selection [Houk et al., 1995] for the basal ganglia. Our model therefore shares part of the view proposed by [Gurney et al., 2001; Prescott, 2001; Redgrave et al., 1999] on the basal ganglia being a centralised behaviour selector. However, our implementation fundamentally differs from theirs in that the role of dopamine (DA) is to signal the error in the prediction of reward and not to be a threshold in the process of elicitation of behaviour.

The view we introduce also connects to former studies on *behaviour selection* in ethological robotics [Ávila García and Cañamero, 2002], which explain behaviour selection as a comparison of behavioural intensities (proportional to the effect on each motivation). We adhere to the view of a motivation driven behaviour selection [Toates and Jensen, 1990] that we also extend to learning. In fact, in a dynamic environment it is likely that only the capacity of re-building the relationships between motivational states, physiological effects and behaviours dramatically facilitates animal's and animat's survival. We argue that the actor-critic algorithm [Sutton and Barto, 1981] ingrains these principles in the context of reward driven decision making and learning while does not impose any formulae to combine stimuli into motivations.

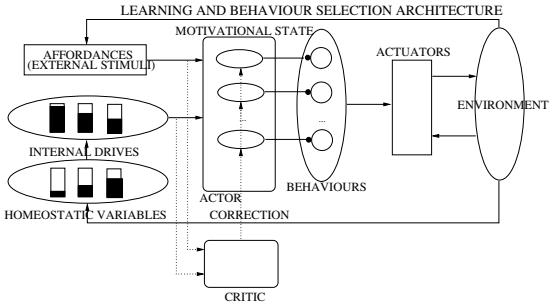


Figure 1: Model Schema.

These ideas are sufficient to motivate an actor-critic based architecture to learn behavioural patterns. However, physiological and anatomical data only is insufficient to propose a sensible criterion for learning to select behaviours. In fact, there are as many possible correct choices for a given a motivational state as criteria. However, some constraints can be extracted from the combination of principles of ethological coherence combined with the rules of convergence of the actor-critic. The final behavioural patters will maximise reward, therefore it is straightforward to relate this principle to Ashby's notion of viability: good actions increase internal physiological stability and therefore are associated to a positive reward, conversely for bad actions [Ashby, 1965]. This single principle should be sufficient to constrain the computations of the motivational state by the actor-critic towards patterns that ensure internal stability.

The actor-critic was designed to this aim. However, the procedure with which stimuli combine to do so entirely depend on how reward is defined. This paper aims at shedding some light in this and at providing a qualitative and quantitative measure of the influence that reward and stimuli have on the learning process and on the apprehended behavioural patterns. The rest of the paper is divided into three sub-sections: a description of the model, an explanation of the experimental setup, and finally some results and conclusions.

2 Learning Behavioural Patterns

The architecture consists of an artificial internal physiology and module for arbitration and learning after biological inspiration. These are introduced next.

2.1 Internal Physiology

The internal physiology is a subset of that described in [Cañamero, 1997]. It consists of a set of homeostatic variables — survival-related variables representing the agent's internal resources —, a set of drives that signal the need to compensate any homeostatic variable, a behaviour repertoire and an arbitration mechanism to resolve conflicts among competing motivations to choose a behaviour. Due to limitations of space it has been impossible to include an analytical description, refer to [Cos-Aguilera et al., 2003] for further information. The model also contains two *hormones*: *satisfaction*, released when there is a successful conclusion of an interaction with an objects and *frustration*, triggered in the converse case.

2.2 Learning and Behaviour Selection

The arbitration mechanism is embedded in the actor-critic algorithm, cf. centre of figure 1, which exhibits separate modules for behaviour selection (actor) and learning (critic). However, both roles are interrelated via the motivational state $\bar{s}(t)$, which consists of the drives $\bar{d}(t)$ and the affordances of the closest object in the neighbourhood of the agent $\bar{a}(t)$, represented as

$$\bar{s}(t) = \{\bar{a}(t), \bar{d}(t)\}. \quad (1)$$

The *actor* performs *behaviour selection* by calculating the likelihood for each behaviour of leading to maximum cumulative reward given the current motivational state $\bar{s}_t$ and choosing the behaviour to execute next b_i according to a winner-take-all policy.

The *critic* estimates the cumulative reward $V(\bar{s}(t))$ resulting from the execution of the behavioural pattern (decided by the actor) leading from the current motivational state to the optimal zone. This is analogous to the Pavlovian learning observed by [Schultz et al., 1993] and extended by [Houk et al., 1995] to the instrumental case. The stimulus for our case consists of the motivational state which the critic relates to a reward via Temporal Difference updates (TD) as expressed by

$$\delta(t) = r(\bar{s}(t-1), b_k) - (V(\bar{s}_{t-1}) + \gamma V(\bar{s}_t)), \quad (2)$$

where $r(\bar{s}(t-1), b_k)$ represents the real reward obtained due to the execution of behaviour b_k. Therefore, the *learning* explicitly consists of *a-posteriori* update of the weights of the estimators that predict the reward ($V(\bar{s}(t-1))$) for the previous motivational state in the case of the critic. Furthermore, the critic also assesses the policy that led to the execution b_k for that motivational state $\bar{s}(t-1)$ by updating the weights of the NN that estimated the policy of the behaviour b_k by a value proportional to δ. The underlying idea is that if the execution of that behaviour was successful that policy should be incentivazed.

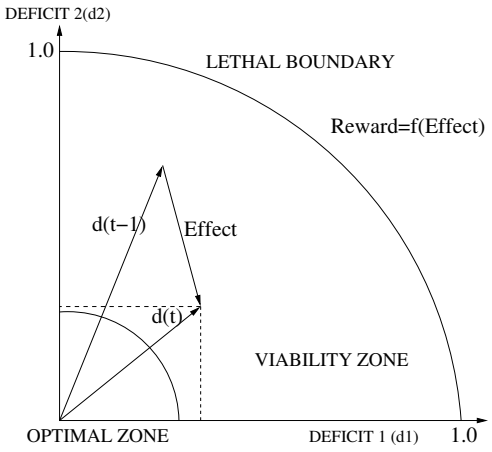


Figure 2: Definition of Reward (case of 2 drives only).

2.3 The Definition of Reward

Reward is interpreted in this approach as the affective interpretation of the physiological effect provoked by the execution of a behaviour. This sense may have arisen because it facilitates survival in a competitive niche. It is straightforward that animals that associate a good feeling to behaviours that improve their physiological state have a larger probability of surviving that the others. This is an enhanced solution to the need of maintaining physiological stability proposed by [Ashby, 1965], which we propose to use as an ethological constraint. Therefore, *reward* can be modelled as an assessment of the physiological effect provoked by a behaviour. Among the infinite formulae that quantify behaviour and the aforementioned constraint we have chosen the following

$$r(t) = \frac{1}{\|\bar{d}_i(t)\|^2} - \frac{1}{\|\bar{d}_i(t-1)\|^2}, \quad (3)$$

where $\bar{d}_i(t)$ and $\bar{d}_i(t-1)$ are the current and the previous physiological states, respectively. This formula relates effects diminishing the deficits to a positive value, coherently with the aforementioned constraint (cf. figure 2).

The hypothesis introduced by formula 3 is vital for several reasons. On the one hand it introduces the sufficient constraints to extend the learning hypothesis of [Schultz et al., 1993] to instrumental learning [Houk et al., 1995], since now the delivery of reward is always mediated by the execution of the appropriate behaviour. On the other, this formula respects basic ethological constraints while does not impose any arithmetic formulae to combine external and internal stimuli to compute the motivational state. The only inherent condition imposed to the behavioural patterns of the algorithm is that these must maximise the reward within the cycle of execution.

3 Experiments

A set of experiments to quantitatively relate the behavioural patterns obtained by testing this model to its internal dynamics in a variety of significant environments have been performed. Their results are introduced in the next section.

3.1 Experimental Setup

For both experiment sets, the robot is placed in two sets of *environments* containing some objects. Their *affordances* have been distributed in such a manner that small objects afford grasping, large ones afford to shelter and all objects afford to be touched. The scenarios have been engineered in order to vary the *availability* and *accessibility* of resources in the environment. In scenario E1 (Motivation Driven) afford every behaviour to be performed. On the contrary each object in scenario E3 (Stimulus Driven) afford a single behaviour to be performed. Scenario E2 (Motivation and stimulus Driven) is a middle case between cases E1 and E3. The robot knows *a-priori* the affordance values of each objects and has to learn the appropriateness of each behaviour to satisfy one need or another.

The metric used to assess the learning mechanism is the **physiological stability**. Two viability indicators are used, namely, *physiological stability* and *overall comfort*. Physiological stability is the average level of satisfaction for all variables, and overall comfort a measure of the homogeneity with which the needs are satisfied. These indicators respond to the formulas 4 and 5, respectively. Similar indicators have been introduced in [Ávila García and Cañamero, 2002], where the lifespan was used to normalise the indicators. This is not necessary in our experimental schema, since simulations have a fixed length.

$$Physiological\ Stability = \frac{1}{N}\sum_{i=0}^{N-1}\hat{d}_i(t) \tag{4}$$

$$Overall\ Comfort = \frac{1}{N}\sum_{i=0}^{N-1}\sigma(\hat{d}_i(t)) \tag{5}$$

The robot navigates at random. Every time an object is encountered, the state is updated by perceiving the set of external (*affordances*) of the object encountered and by reading the instantaneous value of the agent's internal drives (*drives*). The actor calculates then the motivational state (the policy values) and the behaviour whose related motivation exhibits the highest value is selected and executed. Then the object is abandoned, to wander at random until another object is encountered to re-start the cycle of execution.

3.2 Learning Behavioural Patterns

The goal of the first set of experiments is dual. Firstly, it intends to *test the performance of the model to integrate external and internal stimuli to produce behavioural sequences that contribute to the internal physiological stability*. The actor-critic should be able to cope with a diversity of environments by providing appropriate policies to lead the system towards stability. The choice of the three aforementioned environments covers, from a behavioural perspective, the range of adaptation we aim to study.

Secondly, it addresses the study of the *dependence of the learning process on the perception of the agent*. To that aim, experiments have been parametrised after a distortion parameter α. Related to this, $n(t)$ has been added to the perceived affordance values ($a_i^{'}(t)$) as white additive noise ($n(t), m_x = 0, amplitude = \alpha$), cf. equation 6. Sets of 2 simulations have been run for each value of noise, varying their amplitude (α) between 0.0 and 1.0 in increments of 0.2.

$$a_i(t) = a_i^{'}(t)(1-\alpha) + \alpha(n(t) + 0.5) \tag{6}$$

Equation 6 shows the affordance value resulting from the addition of Gaussian noise ($n(t), m_x = 0, amplitude = \alpha$) to its original value ($a_i^{'}(t)$).

he *learning process* is organised in cycles, each commencing by re-setting the homeostatic variables at a random value between 0.0 and 1.0. The agent will then have to make appropriate decisions until the norm of the vector of deficits (drives) is in the optimal zone (cf. figure 2). When this happens, the values are newly reseted to start the following cycle.

Results For all simulations, the length of the cycle of execution decreases overtime a minimum value. Furthermore, the shorter the cycle the smaller the mean of the deficits (physiological balance), cf. fig. 3.

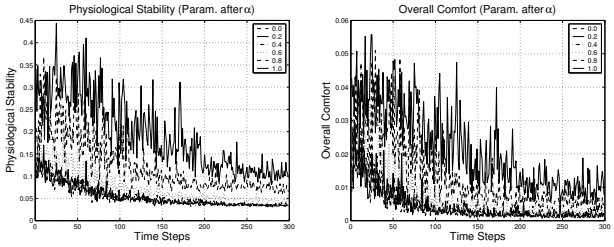


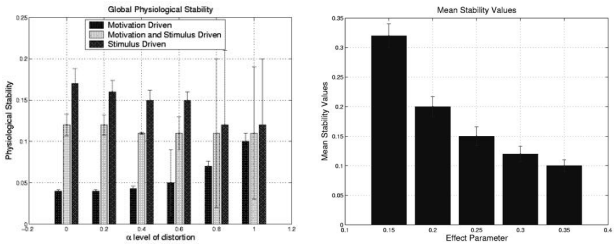
Figure 3: Physiological Parameters along the Simulation for EI, mean and variance, left and right, respectively

However, depending on the environment (cf. picture (a) in figure 4), the distortion (α) may facilitate or not convergence. If affordances are very available (Environment EI) distortion renders convergence more difficult. Interestingly, for incentive driven behaviours (EIII) convergence is facilitated by distortion. This may be explained by the combination of slow decay of the agent's variables and of difficulty in interpreting the external stimulus when distortion is high. For its highest value ($\alpha = 1.0$), the selection will logically be only based on the internal drives, disregarding the external stimulus.

3.3 From Effect to Reward

The goal of the experiment set is to *evaluate the influence of the definition of reward on the basis of the physiological effect*. The amount of effect and its consequent interpretation as reward during the learning process determine not only the pace at learning, but also the quality of the final values for convergence. To this aim, the effect of a behaviour execution on a homeostatic variable has been parametrised between 0.15 and 0.35.

Results The results shown in picture (b) of figure 4 show that this influence of the relationship between effect and reward is vital to determine the final stability of the learnt policy. The larger the effect step size, the shorter the cycle and



(a) Physiological Stability parametrised after α Distortion Values

(b) Physiological Stability parametrised after Effect Values

Figure 4: Physiological Stability for the cases of distorted affordances and of different quantisations of effect, left and right, respectively.

the larger the physiological stability. This has to be considered in the ecological framework that characterises the relationship between the agent, in terms of internal physiology, and the environment in terms of how this affects the agent. The more available are the resources in the environment, the easier it should be to learn policies to satisfy internal needs. Likewise, the larger is the effect with regard to the pace of growth of the deficits, the shorter is the cycle and the lower the mean of the deficits.

4 Discussion and Future Work

The architecture introduced in this paper suggests a simple manner of integrating affordances and internal stimuli for learning behaviour selection in a biologically inspired fashion. It has demonstrated to provide appropriate policies to maintain physiological stability in a variety of scenarios, with different availability and accessibility of resources. This supports the hypothesis for the role of dopamine (DA) in the basal ganglia not only as the error of the prediction of reward for Pavlovian, but also for instrumental learning.

This also highlights that instrumental learning is linked to the definition of reward, which is also related to the physiological stability. Actions contributing to stability should be considered beneficial, conversely for others. Learning radically depends on this. Furthermore, agents can only live within a range of reward definitions relating their internal physiological dynamics and their environment. This restriction seems to be imposed by the need of physiological stability.

Furthermore, the experiments also highlight the fact that the effect of the affordances on the behaviour selection is highly noticeable. In their absence or when they are blurred, the time to learn efficient policies increases, turning the choice of behaviour into a blind selection. However, the scenario and the rhythms of the agent's internal physiology need to be considered to make sense of it, therefore reinforcing the ecological principle.

Based on this, we suggest that animals living in a fast changing environment may exhibit the ability to learn behavioural patterns mostly in a developmental manner and that these are solely assessed via the agent's ecological relationship to its environment.

The mechanism to select the behaviour to execute next is a simple winner-take-all of the policy values, which is only one of the several mechanisms that can explain ethological observations. The choice of what to do next follows often complex ways, which may not necessarily correspond to the aforementioned straight forward mechanisms to maintain the stability of the *internal milieu*. This will be addressed in the near future.

References

[Ashby, 1965] Ashby, W. (1965). *Design for a Brain: The Origin of Adaptive Behaviour*. Chapman & Hall, London.

[Ávila García and Cañamero, 2002] Ávila García, O. and Cañamero, L. (2002). A comparison of behaviour selection architectures using viability indicators. In *Proc. of International Workshop on Biologically-Inspired Robotics: The Legacy of W. Grey Walter*. Bristol HP Labs, UK.

[Cañamero, 1997] Cañamero, L. D. (1997). Modeling motivations and emotions as a basis for intelligent behavior. In Johnson, W. L., editor, *Proceedings of the First International Symposium on Autonomous Agents (Agents'97)*, pages 148–155. New York, NY: ACM Press.

[Cos-Aguilera et al., 2003] Cos-Aguilera, I., Cañamero, L., and Hayes, G. M. (2003). Learning object functionalities in the context of behavior selection. In *Proceedings of the 3rd. Conference Towards Intelligent Mobile Robotics*.

[Dayan and Balleine, 2002] Dayan, P. and Balleine, B. W. (2002). Reward, motivation and reinforcement learning. *Neuron*, 36:285–298.

[Gurney et al., 2001] Gurney, K., Prescott, T., and Redgrave, P. (2001). A computational model of action selection in the basal ganglia. i. a new functional anatomy. *Biological Cybernetics*, 84:401–410.

[Houk et al., 1995] Houk, J. C., Adams, J. L., and Barto, A. G. (1995). Models of information processing in the basal ganglia. In Houk, J. C., Davis, J. L., and G., B. D., editors, *A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement*, A Bradford Book, chapter 13, pages 249–270. MIT Press, 2nd. edition (1998) edition.

[Prescott, 2001] Prescott, T. J. (2001). The evolution of action selection. In Holland, O. and McFarland, D., editors, *The whole iguana*. Cambridge MA: MIT Press.

[Redgrave et al., 1999] Redgrave, P., Prescott, T., and Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89:1009–1023.

[Schultz et al., 1993] Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, (13):900–913.

[Sutton and Barto, 1981] Sutton, R. and Barto, A. (1981). Towards a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 198(88):135–170.

[Toates and Jensen, 1990] Toates, F. and Jensen, P. (1990). Ethological and psychological models of motivation - towards a synthesis. In Meyer, J.-A. and Wilson, S. W., editors, *Proceedings of the First International Converence on Simulation of Adaptive Behaviour*, A Bradford Book., pages 194–205. The MIT Press.

Reinforcement Learning of Stable Trajectory for Quasi-Passive-Dynamic Walking*

Kentarou HITOMI[1], **Tomohiro SHIBATA**[1,2], **Yutaka NAKAMURA**[1], **and Shin ISHII**[1]
[1]Nara Institute of Science and Techcnology
Takayama 8916-5, Ikoma, Nara 630-0192
{kenta-hi,tom,yutak-na,ishii}@is.naist.jp
[2]ATR Computational Neuroscience Laboratories

Abstract

A class of biped locomotion called Passive Dynamic Walking (PDW) has been recognized to be efficient in energy consumption and a key to understand human walking. Although PDW is sensitive to the initial condition and disturbances, some studies of Quasi-PDW, which incorporates supplemental actuators, have been reported to overcome the sensitivity. In this article, we propose a reinforcement learning scheme designed in particular for Quasi-PDW walking. The keys of our approach are a reward function and a learning method of a simple intermittent feedback controller, both of which utilize the robot's passive dynamics as much as possible. They successfully make the action selection problem for walking significantly reduced. Computer simulations show that the parameter in a Quasi-PDW controller is quickly learned after only 180 episodes, and that the obtained controller is robust against sudden perturbations and variations in the slope gradient.

1 Introduction

Biped walking is one of the major research topics in recent humanoid robotics, and many researchers are now interested in Passive Dynamic Walking (PDW) rather than the conventional Zero Moment Point (ZMP) criterion. The ZMP criterion is usually used for planning a desired trajectory to be tracked by a feedback controller, but the continuous control to maintain the trajectory consumes a large amount of power. In contrast, PDW is completely unactuated walking on a gentle downslople [McGeer, 1990]. However, PDW is generally sensitive to the robot's initial posture, speed, and disturbances incurred when a foot touches. To overcome this sensitivity, "Quasi-PDW" [Sugimoto and Osuka, 2003; Takuma *et al*., 2004; Wisse and Frankenhuyzen, 2003] methods, in which some actuators are activated supplementarily to handle disturbances, have been proposed. Because Quasi-PDW is a modification of the PDW, this control method consumes much less power than control methods based on the ZMP criterion. In the previous studies of Quasi-PDW, however, parameters of an actuator had to be tuned based on try-and-error by a designer or on *a priori* knowledge of the robot's dynamics. To act in non-stationary and/or unknown environments, it is necessary for robots that such parameters in a Quasi-PDW controller are adjusted automatically in each environment.

In this article, we propose a reinforcement learning method to train a feedback controller for Quasi-PDW. In our method, we define the reward as becoming large when the robot repeats same motions, i.e., the trajectory of the locomotion is stable. Computer simulation shows that a good controller which realizes a stable quasi-passive walking by a biped robot with knees, can be obtained with a relatively small number of iteration of learning, whereas the controller before learning has poor performance such to allow the biped robot to walk for only a few steps.

In an existing study [Tedrake *et al*., 2004], a stochastic policy gradient reinforcement learning was successfully applied to a controller for Quasi-PDW, but their robot was presumably stable and relatively easy to control because it had large feet whose curvature radius was almost the same as the robot height, but no knees. Furthermore, the reward was set according to the ideal trajectory of the walking motion, which had been recorded when the robot realized a PDW. In contrast, our robot model has closer dynamics to humans where there are smaller feet whose curvature radius is one-fifth of the robot height, and knees. The reward is simply designed so as to produce a stable walking trajectory, without explicitly specifying a desired trajectory. Furthermore, the controller we employ performs feedback control for a short period especially when both feet touch the ground, whereas the existing study above employed continuous feedback control. We believe that our approach is more plausible in the perspective of energy efficiency and understanding of human walking.

2 Approach Overview

Fig. 1 depicts the biped robot model composed of five links connected by three joints: a hip and two knees. The motions of these links are restricted in the sagittal plane. The angle between a foot and the corresponding shank is fixed. Because we intend to explore an appropriate control strategy based on the passive dynamics of the robot in this study, its physical parameters are set referring to the existing biped robots

*This research is supported by JSPS Grant-in-Aid for Scientific Research No. 15300102.

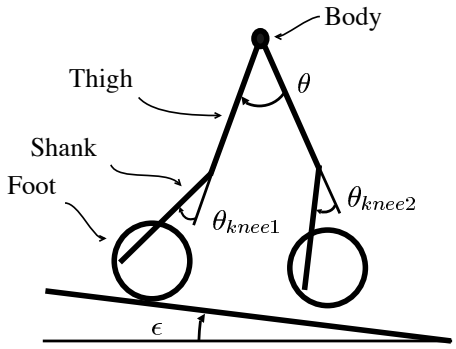


Figure 1: 2D biped robot model

that produced Quasi-PDW [Wisse and Frankenhuyzen, 2003; Takuma *et al*., 2004]. The length and weight of a thigh or shank are 0.33 [m] and 0.5 [kg], respectively. The curvature radius of a foot is 0.132 [m], which is one fifth of the robot's height. The weight of a foot is 0.02 [kg]. The body is a point mass of 4 [kg]. As described in Fig. 1, θ stands for the absolute angle between the two thighs, θ_{knee1} and θ_{knee2} denote the knee angles, and ω denotes the angular velocity of the body around the point at which the stance leg touches the ground. The motion of each knee is restricted within $[0, \pi/4]$ [rad].

Our approach to achieving adaptive controls consists of the following two stages.

1. The two knees are locked, and the initial conditions, θ^s and ω^s, which realize PDW are searched for. These values are used for the initial setting of the robot in the next stage.
2. The two knees are unlocked, and the robot is controlled by a feedback controller with a variable gain parameter. The parameter is modified by reinforcement learning so that the robot keeps stable walking.

These two stages are described in detail in the followings.

2.1 Searching for the initial conditions

In the first stage, we searched for an initial posture, denoted by θ^s and ω^s, which realize PDW on a downslope with a gradient of $\epsilon = 0.03$ [rad]. For simplicity, we fixed $\theta^s = \pi/6$ [rad] and searched a region from 0 to π [rad/sec] by $\pi/180$ [rad/sec], for ω^s that maximizes the walking distance. We found $\omega^s = 58 \times \pi/180$ [rad/sec] was the best value such to allow the robot to walk for seven steps.

2.2 Feedback controller

In light of the design of control signals for the existing Quasi-PDW robots [Wisse and Frankenhuyzen, 2003; Takuma *et al*., 2004], we introduce torque inputs of a rectangular shape applied to the hip joint (cf. Fig. 3). A single rectangular control torque is characterized by a three-dimensional vector $\boldsymbol{\tau}$, that is, τ_{Lag}: the lag time of the torque applying time from the time when either leg is off the ground, τ_{Amp}: the torque amplitude, and τ_{Dur}: the duration of the torque application (Fig. 3(2)). At each time when both legs touch the ground, the controller observes the state of the robot, i.e., θ, and then outputs a control torque $\boldsymbol{\tau}$ applied to the biped robot. $\boldsymbol{\tau}$ is assumed to be distributed as a Gaussian noise vector (see section 3.2) whose means are given by

$$\bar{\tau}_{Lag} = w_{Lag,\theta}\left(\theta - \theta^d\right) + w_{Lag,b} \quad (1)$$

$$\bar{\tau}_{Amp} = w_{Amp,\theta}\left(\theta - \theta^d\right) + w_{Amp,b} \quad (2)$$

$$\bar{\tau}_{Dur} = w_{Dur,\theta}\left(\theta - \theta^d\right) + w_{Dur,b}, \quad (3)$$

where θ^d is supposed to be the desired value of θ and $w_{Lag,b}$, $w_{Amp,b}$, $w_{Dur,b}$ are bias terms, respectively. The value of θ^d is unknown, however, so we let the robot to learn θ^d by expressing $\theta^d = \theta^s + w^d$ and adjusting the value of w^d. Here, θ^s is used as the initial value of θ^d. Equations(1) $\sim$ (3) are then expressed as

$$\bar{\boldsymbol{\tau}} = \boldsymbol{w}_\theta\left(\theta - \theta^s\right) + \left(-\boldsymbol{w}_\theta w^d + \boldsymbol{w}_b\right), \quad (4)$$

where $\boldsymbol{w}_\theta$,$\boldsymbol{w}_b$ and w^d are the parameters that should be adjusted by reinforcement learning, and $\boldsymbol{w}_\theta$ and $\boldsymbol{w}_b$ are

$$\boldsymbol{w}_\theta = (w_{Lag,\theta}, w_{Amp,\theta}, w_{Dur,\theta})$$
$$\boldsymbol{w}_b = (w_{Lag,b}, w_{Amp,b}, w_{Dur,b}).$$

Because the learning algorithm requires the differentials of equation (4) with respect to the parameters, we linearized this equation as

$$\begin{aligned}\boldsymbol{w}'_b &= \left(w'_{Lag,b}, w'_{Amp,b}, w'_{Dur,b}\right)^T \\ &= -\boldsymbol{w}_\theta w^d + \boldsymbol{w}_b\end{aligned}$$

$$\bar{\boldsymbol{\tau}} = \boldsymbol{w}_\theta\left(\theta - \theta^s\right) + \boldsymbol{w}'_b. \quad (5)$$

The six-dimensional vector:

$$\begin{aligned}\boldsymbol{W} \;=\; & (w_{Lag,\theta}, w'_{Lag,b}, \\ & w_{Amp,\theta}, w'_{Amp,b}, w_{Dur,\theta}, w'_{Dur,b})\end{aligned}$$

was adjusted by an on-line reinforcement learning scheme, which is described in section 3. Since we assume no *a priori* knowledge of the feedback control, $\boldsymbol{W}$ was set at $\mathbf{0}$ as its initial value.

3 Learning a Feedback Controller

3.1 Policy gradient reinforcement learning

In this study, we employ a stochastic policy gradient method [Kimura and Kobayashi, 1998b; 1998a] in the reinforcement learning for the controller's parameter $\boldsymbol{W}$. The robot is regarded as a discrete dynamical system whose discrete time elapses when either foot touches the ground, i.e., when the robot takes a signal step. The state variable of the robot is given by θ_n, where n counts the number of steps. For each state θ_n, the controller provides a control signal $\boldsymbol{\tau}$ according to a probabilistic policy $\pi(\boldsymbol{\tau} \mid \theta_n)$. At the next step, the controller observes a new state θ_{n+1} and is assumed to receive a reward signal r_n. Based on these signals, a temporal-difference (TD) error δ is calculated by

$$\delta = \{r_n + \gamma V(\theta_{n+1})\} - V(\theta_n), \quad (6)$$

where $\gamma(0 \leq \gamma \leq 1)$ is the discount rate. V denotes the state value function and is trained by the following TD(0)-learning:

$$V(\theta_n) \leftarrow V(\theta_n) + \alpha\delta, \tag{7}$$

where α is the learning rate. The policy parameter $\boldsymbol{W}$ is updated as

$$\boldsymbol{e} = \frac{\partial}{\partial \boldsymbol{W}} \ln\left(\pi\left(\boldsymbol{\tau} \mid \theta, \boldsymbol{W}\right)\right)\Big|_{\boldsymbol{\tau}=\boldsymbol{\tau}_n, \theta=\theta_n, \boldsymbol{W}=\boldsymbol{W}_n} \tag{8}$$

$$\boldsymbol{D} \leftarrow \boldsymbol{e} + \beta \boldsymbol{D} \tag{9}$$

$$\boldsymbol{W}_{n+1} = \boldsymbol{W}_n + \alpha_p \delta \boldsymbol{D}, \tag{10}$$

where $\boldsymbol{e}$ is the eligibility and $\boldsymbol{D}$ is the eligibility trace. $\beta(0 \leq \beta \leq 1)$ is the diffusion rate of the eligibility trace and α_p is the learning rate of the policy parameter. After policy parameter $\boldsymbol{W}_t$ is updated into $\boldsymbol{W}_{t+1}$, the controller emits a new control signal according the new policy $\pi\left(\boldsymbol{\tau} \mid \theta_{n+1}, \boldsymbol{W}_{n+1}\right)$. Such a concurrent on-line learning of the state value function and the policy parameter is executed until the robot tumbles (we call this period an episode), and the reinforcement learning proceeds by repeating such episodes.

3.2 Simulation setup

In this study, we try to obtain an appropriate controller for the hip joint, whereas the knee joints are controlled by a predetermined simple deterministic controller. The stochastic policy is defined as a normal distribution:

$$\pi(\boldsymbol{\tau} \mid s, \boldsymbol{W}) = \frac{1}{(2\pi)^{3/2} \mid \boldsymbol{\Sigma} \mid^{1/2}} \times \exp\left\{-\frac{1}{2}(\boldsymbol{\tau} - \bar{\boldsymbol{\tau}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\tau} - \bar{\boldsymbol{\tau}})\right\}, \tag{11}$$

where the mean $\bar{\boldsymbol{\tau}}$ is determined by the feedback control rules (equations (1)(2)(3)) and the covariance $\boldsymbol{\Sigma}$ is given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} {\sigma_{Lag}}^2 & 0 & 0 \\ 0 & {\sigma_{Amp}}^2 & 0 \\ 0 & 0 & {\sigma_{Dur}}^2 \end{pmatrix}, \tag{12}$$

where σ_{Lag}, σ_{Amp} and σ_{Dur} are set at $0.001, 0.3$ and 0.05, respectively. We assume each component of $\boldsymbol{\tau}$ is 0 or positive, and if it takes a negative value probabilistically it is calculated again, similarly in the previous study [Kimura *et al.*, 2003].

The reward function is set up as follows. If a robot walks stably, ω_n and θ_n should repeat similar values over steps. Furthermore, the robot should take no step in the same place, i.e., θ_{n+1} needs to be large enough.

To satisfy above requirements, we define the reward function as

$$r_n = \theta_{n+1} \exp\left(- \mid \theta_{n+1} - \theta_n \mid\right). \tag{13}$$

Fig. 2 shows the landscape of this reward function.

The knees are controlled by a simple control scheme described below (cf. Fig. 3) so that each leg in the swing phase does not contact to the ground. A torque (1 [Nm]) is applied to the knee joint of the swing leg in order to flex the knee, from the moment that a torque is applied to the hip joint of the swing leg to make this leg go forward, until the foot of the swing leg goes ahead of that of the stance leg. Then,

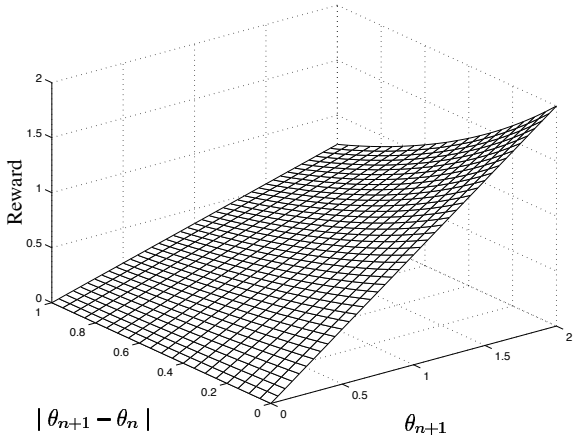


Figure 2: Landscape of the reward function

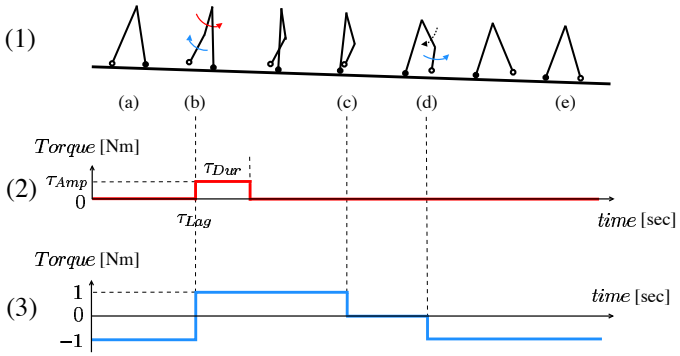


Figure 3: Torque applied to the hip joint and the knee joint. (1) Motions of the swing leg during a single step. (2) Torque applied to the hip joint. (3) Torque applied to the swing leg's knee joint. (a) A single step starts when both feet touch the ground. (b) After τ_{Leg} [sec], the robot begins to bend the swing leg's knee by applying a positive torque. (c) The torque to the knee is removed when the foot of the swing leg goes ahead of that of the stance leg. (d) When the thigh of the swing leg turns into the swing down period from the swing up period, a negative torque is applied in order to extend the swing leg. (e) The swing leg touches down and becomes the stance leg.

the torque is removed so that the swing leg is swang down according to its passive dynamics. After the swing leg turns into the swing down period from the swing up period, a torque of -1 [Nm] is applied in order to make the leg extend; this control is continued until the leg touches the ground and then becomes the swing leg again.

The value function is represented by a table over grid cells in the state space, and the value for each grid cell is updated by equation (7). In this study, we prepared 10 grid cells; the center of the fifth cell on each coordinate was θ^d (Fig. 4), and the grid covered the whole state space, by assigning the 0-th cell on each coordinate to the range $\theta < 0$. We used $\alpha = 0.5, \alpha_p = 0.01$, and $\beta = \gamma = 0.95$.

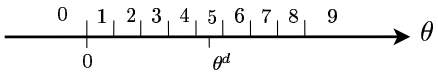


Figure 4: Discretization of the state space

In this study, we used a 3D dynamics simulator, Open Dynamics Engine [ODE,]. In simulation experiments, motions of the robot were restricted in the sagittal plane by configuring a symmetric robot model with nine links (Fig. 5). It should be noted this nine-links robot has equivalent dynamics to the five-links model (Fig. 1), under the motion restriction in the sagittal plane; this nine-links model was also adopted in Wisse [Wisse and Frankenhuyzen, 2003] and Takuma et al. [Takuma *et al*., 2004].

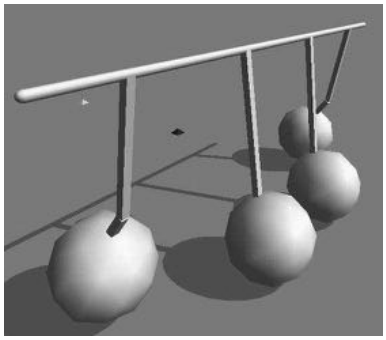


Figure 5: Dynamics simulation of the nine-links model with ODE

4 Simulation Results

Although the physical parameters of our robot were set referring to the existing Quasi-PDW robots, our robot with unlocked knees was not able to produce stable walking by itself. Then, this section describes the way to train the controller.

4.1 Passive walking without learning

First, we examined whether the robot with unlocked knees was able to produce stable walking on a downslope with $\epsilon = 0.03$ [rad], when it received no controls to the hip joint. The unlocked knees were controlled in the same manner as that described in section 3.2. Initial conditions were set at $\theta_0 = \theta^s$ [rad] and $\omega_0 = \omega^s$ [rad/sec], which are the same as those in the knee-locked model that performed seven steps walking. As Fig. 6 shows, the robot with unlocked knees walked for 80 cm and then fell down. The robot could not walk passively when the knees were unlocked but controlled by a simple heuristic controller.

4.2 Learning a feedback controller

The experiment in section 4.1 showed that the robot with unlocked knees was not able to produce stable walking without any control to the hip joint, even when starting from good

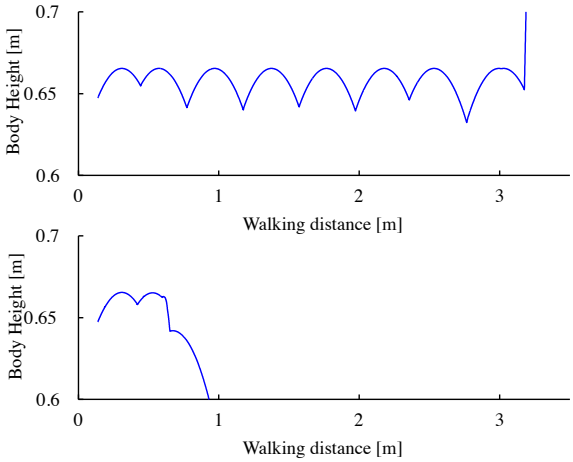


Figure 6: Effect of knee existence in PDW. Upper panel shows a trajectory of the knee-locked model, while lower panel is for when knees were unlocked but controlled by a simple controller. In both panels, the hip was not controlled.

initial conditions θ^s and ω^s. Then, in this section, we applied on-line reinforcement learning to the automatic tuning of the parameter $\boldsymbol{W}$ in the feedback control rules, equations (1), (2), and (3). At the onset of each episode, the robot was set to its initial conditions $\theta_0 = \theta^s, \omega_0 = \omega^s$, and the episode was terminated either when the robot walked for 50 steps or fell down. When the height of the robot's 'Body' became smaller than 80% of its maximum height, it was regarded as a failure episode (falling down). Reinforcement learning was continued by repeating such episodes.

Fig. 7 shows the moving averages for ± 20 episodes of walking steps (upper) and cumulative reward (lower), achieved by the robot. The steps increased after about 80 episodes, and went up to near 40 steps after about 180 episodes. In the early learning stage, the cumulative reward and walked distance were small though the robot walked for more than 10 steps, indicating the robot was walking stumbling with small strides. Using the deterministic feedback controller with the parameter $\boldsymbol{W}$ after 180 training episodes, the robot could walk for more than 50 steps (Fig. 8). The parameter at this time was $\boldsymbol{W} = (-4.725 \times 10^{-4},\ 9.558 \times 10^{-4},\ -4.809 \times 10^{-1},\ 7.168 \times 10^{-1},\ -9.481 \times 10^{-3},\ 1.217 \times 10^{-1})$; the negative sign of all feedback coefficients which form $\boldsymbol{w}_\theta$ in equation (5) implies that $\boldsymbol{W}$ had grown to represent an appropriate feedback gain.

4.3 Robustness against disturbances

To see the robustness of the acquired Quasi-PDW against disturbances, we applied impulsive torque inputs to the hip joint during walking. Fig. 9 shows the time-series of θ_n in the same condition as Fig. 8, except that impulsive torque inputs were applied as disturbances at the time points with the arrows. Each disturbance torque was 1 [Nm] and was applied so as to pull the swing leg backward for 0.1 [sec] when 0.4 [sec]

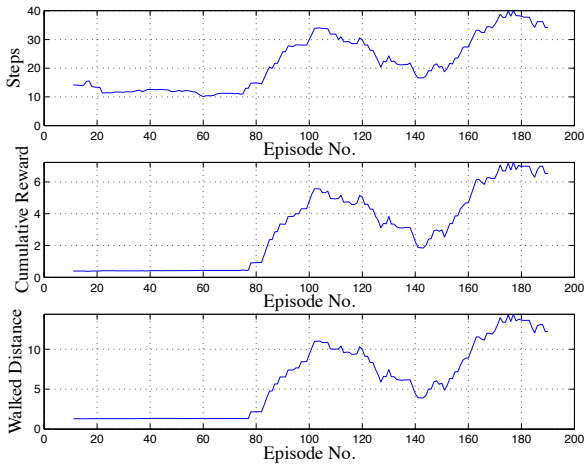


Figure 7: Moving averages of steps and cumulative reward. After about 180 episodes, the robot became to walk for nearly 40 steps and for over 10 [m] in average. Each panel shows the moving average for the period: $-20 \sim +20$.

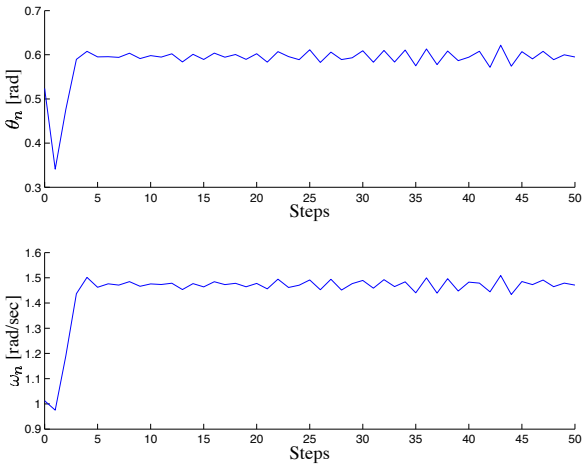


Figure 8: Values of θ_n and ω_n during the walking for 50 steps. The deterministic feedback controller with $\boldsymbol{W}$ acquired after 180 learning episodes was used. Robot walked for more than 50 steps.

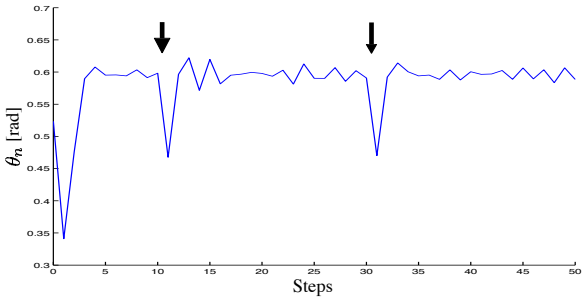


Figure 9: Perturbation of θ against impulsive disturbances (torque). Each torque of 1 [Nm] was applied to the hip joint for 0.1 [sec] during the 10-th, 30-th steps (arrowed) so as to pull the swing leg backward.

elapsed after the swing leg got off the ground. As this figure shows, θ_n recovered to fall into the stable limit cycle within a few steps after disturbances, implying that the attractor of the acquired PDW is fairly robust to noises from the environment.

4.4 Walking on different slopes

Next, we let the robot with the control parameter after 180 training episodes walk on downslopes with various gradients. Fig. 10 shows the results for $\epsilon = 0.02 \sim 0.06$ [rad]. The robot was able to walk for more than 50 steps on downslopes with $\epsilon = 0.02 \sim 0.05$ [rad], and 10 steps with 0.06 [rad]; the feedback controller acquired through the on-line reinforcement learning was robust against the change (in the gradient) of the environment with which θ became larger as the environment got harder.

5 Discussion

In this article, we proposed an on-line reinforcement learning scheme for a feedback controller in order to realize Quasi-PDW which is suitable for locomotive robots in the perspective of low energy consumption and good correspondence to human walking. Our scheme was successful in making the robot with knees produce stable walking after 180 training episodes, despite of the simple feedback controller. Our method acquired a good feedback controller that allows the robot to be entrained to a stable limit cycle based on the passivity of the robot's dynamics.

Our learning scheme consisted of two stages, as described in section 2. After roughly searching in the first stage for an initial angular velocity with which the robot with locked knees walked for several steps, reinforcement learning was applied to the robot with unlocked knees, starting from the initial condition obtained in the first stage. This two-stages learning can be regarded as a developmental progression found at least in humans [Newell and Vaillancourt, 2001; Bernstein, 1968] which increases the degree of freedoms as the learning proceeds; after a primitive control is achieved for a system with a low dimensionality, the dimensionality is gradually released to realize more complex and smooth movements by the high-dimensional system. Furthermore, animals seem to employ different controllers in the initiation phase and in the maintenance phase for effective motor controls [Pahapill and Lozano, 2000]; e.g., it has been known that three steps in average are required to initiate stationary walking in humans [Miller and Verstraete, 1996]. We consider the first stage of our approach corresponds to the initiation stage above.

As another reason for our successful result, our adaptive feedback controller is trained by the on-line reinforcement learning as to apply intermittent energy for maintaining stable PDW. Although the feedback controller itself was simple, the simulation experiments on downslopes with various gradients and through addition impulsive disturbances have shown that the stochastic policy gradient method with the re-

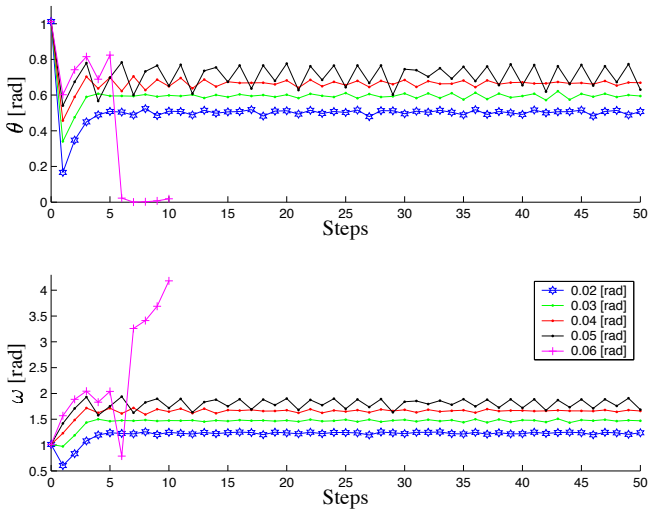


Figure 10: Values of θ_n and ω_n on downslopes with various gradients. Robot could walk more than 50 steps on downslopes with $\epsilon = 0.02 \sim 0.05$ [rad]. On downslope with 0.06 [rad], the strides in the latter five steps were almost 0.

ward given to continue rhythmic walking steps contributed to making the PDW by the robot robust against noises in the environment. This intermittent control was inspired by the studies on the measurement of human EMG [Basmajian, 1976] and on Quasi-PDW conducted by Collins et al.[Collins *et al.*, 2005] or by Takuma [Takuma *et al.*, 2004]. To develop an energy-efficient control method of robots, considerable care about the passivity of the robot should be taken, as Collins suggested. Furthermore, the dynamics of robots with many degrees of freedom is generally a nonlinear continuous system , and thus the action selection for controlling such a system is usually very difficult. Our approach successfully realized rapid learning by introducing the policy that emits intermittent control signals and a reward function encouraging stable motion, both of which utilized the passivity of the robot. Our learning scheme is not restricted to locomotion, since the computational problem and the importance of passivity are both general, although what kind of controllers should be activated or switched when and how are remained as significant problems.

As a future work, we will devise a method to produce stable walking on a level ground. In addition, we will conduct experiments with a real biped robot.

Acknowledgment

We thank Professor Hosoda and Mr. Takuma at Graduate School of Engineering, Osaka University, for kindly giving us information about Passive Dynamic Walking and their biped robot.

References

[Basmajian, 1976] V. Basmajian, J. *The human bicycle*, volume 5-A. University Park Press, Baltimore, 1976.

[Bernstein, 1968] N. Bernstein. *The coordination and regulation of movements*. Pergamon, 1968.

[Collins *et al.*, 2005] S. Collins, A. Ruina, R. Tedrake, and M. Wisse. Efficient bipedal robots based on passive-dynamic walkers. *Science*, 307, 2005.

[Kimura and Kobayashi, 1998a] H. Kimura and S. Kobayashi. An analysis of actor/critic algorithms using eligibility traces: Reinforcement learning with imperfect value function. In *15th International Conference on Machine Learning*, pages 278–286, 1998.

[Kimura and Kobayashi, 1998b] H. Kimura and S. Kobayashi. Reinforcement learning for continuous action using stochastic gradient ascent. *Intelligent Autonomomus Systems*, pages 288–295, 1998.

[Kimura *et al.*, 2003] H. Kimura, T. Aramaki, and S. Kobayashi. A policy representation using weighted multiple normal distribution. *Journal of the Japanese Society for Artificial Intelligence*, 18(6):316–324, 2003.

[McGeer, 1990] T. McGeer. Passive dynamics walking. *The International Journal of Robotics Research*, 9(2):62–82, 1990.

[Miller and Verstraete, 1996] A. Miller, C. and C. Verstraete, M. Determination of the step duration of gait initiation using a mechanical energy analysis. *Journal of Biomechanics*, 29(9):1195–1199, 1996.

[Newell and Vaillancourt, 2001] K. M. Newell and D. E. Vaillancourt. Dimensional change in motor learning. *Human Movement Science*, 2001.

[ODE,] ODE. http://ode.org/.

[Pahapill and Lozano, 2000] P. A. Pahapill and A. M. Lozano. The pedunculopontine nucleus and parkinson's desease. *Brain*, 123, 2000.

[Sugimoto and Osuka, 2003] Y. Sugimoto and K. Osuka. Motion generate and control of quasi-passive-dynamic-walking based on the concept of delayed feedback control. In *Proceedings of 2nd International Symposium on Adaptive Motion of Animals and Machines*, 2003.

[Takuma *et al.*, 2004] Takashi Takuma, Seigo Nakajima, Koh Hosoda, and Minoru Asada. Design of self-contained biped walker with pneumatic actuators. In *SICE Annual Conference*, 2004.

[Tedrake *et al.*, 2004] Russ Tedrake, Teresa Weirui Zhang, and H. Sebastian Seung. Stochastic policy gradient reinforcement learning on a simple 3d biped. In *Proceedings of the IEEE International Conference on Intelleigent Robots and Systems (IROS)*, 2004.

[Wisse and Frankenhuyzen, 2003] M. Wisse and J. Frankenhuyzen. Design and construction of mike; a 2d autonomous biped based on passive dynamic walking. In *2nd InternationalSymposium on Adaptive Motioon of Animals and Machines*, 2003.

Hierarchical reactive planning: Where is its limit?*

Cyril Brom
Charles University in Prague
Faculty of Mathematics and Physics
Malostranské nám. 2/25, Prague, Czech Republic
brom@ksvi.mff.cuni.cz

Abstract

Hierarchical reactive methods are very popular in the field of controlling complex artificial intelligent agents. In this paper, we argue that they cannot cope with human-like behaviour. We present a detailed analysis of behaviour of a relatively simple human-like artificial agent, an artificial gardener, whose action selection model is based on hierarchical reactive planning. It is shown that although the agent has no troubles with "survival" in a complex and dynamic environment, its behaviour is not believable in some situations. However, instead of rejecting the judged methodology, we propose how to extend it using certain features of other action selection models.

1 Introduction

There are two main characteristics that make hierarchical reactive methods of action selection popular in the field of controlling complex artificial intelligent agents. The first is a top-down recursive decomposition of top-level agents' behaviours to sub-behaviours or sequences of simple actions. This decomposition eases the design. The second is that an agent's decision procedure focuses attention only to most relevant goals, while ignoring the rest of them temporarily. All reactive methods allow for quick switching between tasks according to the changing environment and agent's internal drives. Consequently, reactive hierarchies reduce combinatorial complexity of control and still can cope with large, unpredictable, real-time environments.

We are working on a research and educational toolkit for prototyping human-like artificial agents, i.e. the agents with the objective to *imitate* behaviour of humans (*h-agents* in the following text) [Brom *et al.*, 2005]. One of our motivations on this research effort is to find an appropriate methodology for controlling h-agents. The methodology must allow an easy behavioural design and must be computationally effective. Thanks to the aforementioned advantages of reactive hierarchies, we turned our attention to this branch of methods. So far, we have prototyped several h-agents "living" in a family house utilizing reactive hierarchies in order to ascertain their applicability.

As was expected, partially owing to Bryson's analysis [2000], our h-agents had no troubles with "survival", that means with satisfying own needs. However, it was not always straightforward to design their behaviour so that it was *believable* enough. Consequently, h-agents did not behave naturally in some situations—i.e., they would not pass Turing test. Because the h-agents otherwise performed well, we aimed at isolating problems and extending pure hierarchical reactive approach, instead of rejecting it.

In this paper, we present observed limitations on believability and suggest how to overcome them. We present the results in a case-study example of an artificial gardener, whose behaviour is structured by so-called *simple hierarchical reactive planning* (S-HRP). For features that could extend this model, we seek both inside and outside of hierarchical reactive family.

In Section 2, we briefly introduce our toolkit and detail our motivation. Then, we describe morning tasks of a "natural gardener". This story represents desired behaviour. In Section 4, we formalize the S-HRP method and describe behaviour of the artificial gardener. In Section 5, we present the results, together with suggestions on extensions of S-HRP. At the end, we discuss applicability of the extended S-HRP considering related action selection models.

2 Motivation: Project Ents

Simulations of artificial humans are becoming increasingly more popular both in the academic and industrial domains. Typical applications include computer games, virtual storytelling, entertainment applications, military simulations and behavioural modelling (e.g. [Prendinger *et al.*, 2004]).

From the technical point of view, each artificial human is viewed as an autonomous intelligent agent [Wooldridge, 2002] that carries out a diverse set of goals in a large dynamic environment with the objective to simulate *believable* behaviour of humans; this agent is a so-called *h-agent*. One of the key issues in this research field is design of a mind of h-agents (i.e., a memory and a procedure that decides what to do next—an action selection algorithm).

Although various theoretical solutions of this issue have been proposed so far (e.g. [Newell, 1990]), and a lot of

* This research was partially supported by the Program "Information Society" under project 1ET100300517.

individual applications using h-agents and languages for their programming exist, it is hard to find any complex toolkit that would couple an artificial environment similar to natural world, a neat graphical user interface and a language for prototyping h-agents' minds by means of various different techniques. Such a toolkit would simplify development of h-agents, enable verifying theories and it could serve as an educational tool for students.

The project Ents [Brom *et al.*, 2005] is a first generation of a toolkit that addresses these issues. It provides:

- A customizable artificial environment similar to natural world, which allows the h-agent to carry out complex human-like tasks (among others eating, sleeping and going to toilet).
- E language, which that enables modelling of h-agents using various different techniques (including for example hierarchical reactive planning, hierarchical finite state machines, and classical planning).
- A linguistic module, which enables talking to h-agents (in Czech language).
- The tool allows for interaction between the h-agents, and between h-agent and a user-agent.

Nevertheless, a toolkit is not just a programming vehicle and a bundle of debugging tools. It is also a design methodology. Therefore, we are focused not only on how to program h-agents, but also on how to *design* their behaviour *simply*. We are now in the phase of evaluation of the first generation of the toolkit and of various models of action selection, while specifying requirements on the toolkit's second generation. Hence, we evaluate, whether hierarchical reactive planning is the methodology appropriate for h-agents domain; and that is the topic of this paper.

The artificial gardener, whose behaviour is observed in the case-study, is prototyped in the project Ents. A model of a "family-house" is used. A screenshot from the simulation is depicted in Fig. 1. More information on the project is available at: `http://ckl.ms.mff.cuni.cz/~bojar/enti/`.

3 The challenge: natural behaviour

This section describes a story of a typical human that spends his morning gardening. Behaviour of the artificial gardener is modelled according to this "natural model" and the course of resulting behaviour is compared to it. In what follows, the artificial gardener will be denoted as the *a-gardener* and its human model as the *n-gardener* (we will use masculine for the n-gardener, neuter for the a-gardener and feminine for both an artificial and a natural neighbour).

Behaviour is observed from the moment the gardeners are going to the garden to the moment they leave it. Two tasks are intended: watering and weeding. The scenario follows:

Because n-gardener knows that a garden hose is punctured, he decides to water by a can. He goes to a chamber for tools. Because he is intended to weeding afterwards, he does not find only a can, but also a bucket, a weeder and a little scoop. Then he takes all of that (the weeder and the little scoop in the bucket) and goes to the garden. He whistles from time to time for joy.

Figure 1: The GUI of the toolkit Ents. From the left: the user-agent and the artificial gardener.

In the course of watering, two events happen:

1. *A neighbour comes and asks him for a can. He promises her to bring the can after he finishes the watering.*
2. *Nature calls. He puts down the can and goes to the toilet. Then he returns and continues with the task.*

When he finishes watering, he goes to lend the can to the neighbour. Then he starts weeding.

In the course of weeding, following event happen:

3. *The neighbour returns the can. The n-gardener leaves the can as she has put it down.*

After he completes weeding, he puts all the tools into the chamber. Then he goes to eat to the dining room.

4 A-gardener: the action selection model

In this section we explain the action selection model of the a-gardener—*simple hierarchical reactive planning*, the S-HRP, and describe behaviour of the a-gardener. We remark that SHRP resembles the planning method of Bryson [2001].

4.1 Simple hierarchical reactive planning

S-HRP is a top-down, reactive method. The former means that the overall behaviour is decomposed into specific goals, which are recursively decomposed into smaller subgoals, until atomic actions are reached. The latter means that the next action an agent has to perform is not selected from a plan generated before an execution starts, but it is computed instantly by means of context-based triggers that continuously monitor an environment or the agent's internal drives. Reactive planners do not "look ahead"; instead, they compute just the next act in every instant. In S-HRP, the problem of what to do next is reduced to switching among sets of triggers associated with some subgoals according to changing circumstances.

S-HRP provides four behavioural structures: atomic actions, processes, top-level goals and sequences.

- *An atomic action* (a-action) is the primitive operation an h-agent can do. E.g. `aStep(hAgentID,placeID)`.
- *A sequence* is a simple sequence of a-actions or processes, e.g. $<a_1, a_2, p_1, a_3, p_2>$ (*p* denotes a process, *a* denotes an action).
- *A process* is a set of *process-steps* (p-steps), which are quadruples *<p, r, c, a>*, where *p* is a priority local to the process (such that p-steps of one process are total-ordered by their priorities), *r* is a releaser, *c* is an optional context and *a* is an action. Releasers and contexts are boolean conditions, an action can be an a-action, a subprocess or a sequence. In the following we will write directly "priorities of releasers" instead of "priorities of p-steps".
- *A top-level goal* is quadruple *<pr*, *r*, *c*, *f>*, where *pr* is a process associated with achieving the goal, *r* and *c* are releaser and context respectively, and *f* is a function of time that serves as a floating priority of the goal.

Subprocesses are nested under each top-level goal in a tree-like hierarchy. Leaves represent a-actions. The behaviour of one agent is represented by a set of such behavioural structures and this set is called a *betree* (it comes from a "behavioural tree"). The betree is always provided in advance by a behavioural programmer/designer and it is not further modified when an h-agent is running (in S-HRP).

All p-steps and top-level goals of the betree are either active or preactive or inactive or sleeping, all a-actions and processes are either executed or not-executed.

At every instant, at least on of top-level goals' releasers must hold. The highest priority goal with the holding releaser is *active*, the other goals with holding releasers are *preactive*; the rest is *inactive*.

A sequence, which is a child of an active node (i.e. a p-step), is *executed*; other sequences are *not-executed*. A process or an a-action, which is a child of an active node (i.e. a p-step or a top-level goal), is *executed*. Exactly one process or exactly one a-action from an executed sequence is also *executed* (the one just being performed). The other processes and the a-actions are *not-executed*.

Each executed process is associated with several p-steps. At every instant, at least one of their releasers must fire. The p-step (under an executed process) with holding releaser and with the highest priority is called *active*. Its siblings with holding releasers (and lower priority) are called *preactive*. Its siblings without a holding releaser (both with higher and lower priority) are called *inactive*. All p-steps under a not-executed process are *sleeping*. In the following, we will often write directly "active/ inactive/... releasers" instead of "active/inactive/... p-steps/goals".

An example of a betree is depicted in Figure 2. Figure 3 shows the action selection algorithm. This algorithm is performed in every time-step by a control unit of an h-agent. When the simulation starts, all nodes of the given betree are marked as not-executed, or sleeping, except for top-level goals, which are inactive.

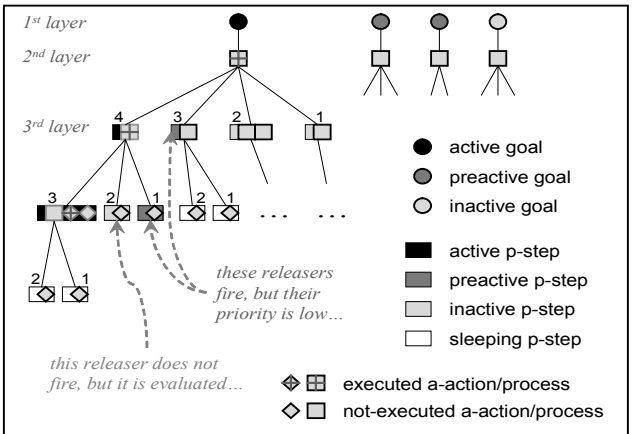


Fig. 2: An example of a betree. The priorities are written above the p-steps.

When the algorithm starts, it evaluates all releasers of top-level goals and identifies an executed process (SELECT-GOAL). Then it recursively finds in a breath-first manner all active, preactive and inactive releasers in the upper layer of the subtree of the executed element (A-S). The active releaser in a leaf triggers an a-action and finishes the evaluation (26). A non-leaf active releaser expands the evaluation

```
SELECT-GOAL( betree ):
(1)  eval ← all releasers of top-level goals of betree  % they "never sleep"
(2)  evaluated ← evaluate eval
(3)  set active/preactive/inactive top-level goals based on evaluated
(4)  set executed/not-executed processes of top-level goals
(5)  p-steps ← all p-steps of executed process
(6)  A-S( third layer of betree, p-steps, betree ) % 3rd layer – see Fig. 2

A-S( layer, p-steps, betree ):
(7)  releasers ← all releasers of p-steps
(8)  eval ← evaluate releasers
(9)  set active/preactive/inactive p-steps from p-steps based on eval
(10) set all other nodes % i.e. p-steps % in the layer as sleeping
(11) act ← the action of the active p-step of p-steps
(12) if act differs from previously executed action then
(13)    set previously executed flag as not-executed
(14) if act is "a-action" or "process" then
(15)    EXEC( act, layer, betree )
(16) else if act is "sequence" then
(17)    if act is not executed then set act as executed
(18)    if act already contains executed element then
(19)       set this element as not-executed        % element = process or
(20)    if this element is not the last element of act then      % a-action
(21)       EXEC( the next element of act, layer, betree )
(22)    otherwise    % restart the sequence:
(23)       EXEC( the first element of act, layer, betree )

EXEC( act, layer, betree ):
(24) set act as executed
(25) if act is "a-action" then
(26)    execute( Act )
(27) otherwise % act is now a "process"
(28)    A-S( next layer of betree, all p-steps of act, betree )
```

Fig. 3: The S-HRP evaluation algorithm.

to the lower layer of the betree (28). Because there is exactly one active releaser in each layer, there is also exactly one active p-step in each layer. Subsequently, exactly one a-action can be performed in the given time step. This a-action can be performed several times, until another leaf releaser becomes active. An agent's attention is switched to another branch of the betree, i.e. to another subtask, when previously preactive or inactive releaser is activated. That happens either when an executed process is finished (i.e. its releaser holds not more), or when external/internal circumstances are changed. As exactly one branch of the betree can be executed, there is no parallelism in S-HRP.

The asymptotic complexity of the S-HRP-evaluation is $O(p.l.a)$, where l is the depth of the betree, p an average number of p-steps of one executed process and a a constant that limits the time for evaluation of one releaser; provided that releasers are re-evaluated in every time step. (Complexity can be reduced significantly by utilizing a variant of RETE algorithm [Forgy, 1982].)

The purpose of a context is to interrupt currently executed action, even if the releaser holds. As contexts are typically conjunctions of releasers with higher priorities, they are omitted from the description of the algorithm for the sake of simplicity (it is possible to express for example timeouts or numbers of retries by means of them). Scheduling of top-level goals is enabled by their floating priorities.

The four elements of S-HRP are similar to the elements of Byson's POSH action selection plans [2001] (primitive actions, action patterns, competences and drive collections). The whole S-HRP-betree resembles to AND-OR tree with only AND branches, which are used for example in total-order simple task network planning [Ghallab *et al.*, 2004].

4.2 Behaviour of the artificial gardener

The behavioural structure of the a-gardener is depicted in Figure 5. For simplicity, library functions like searching for an object or drinking are not detailed. Figure 4 shows the priorities of top-level goals and the whole course of the behaviour. The behaviour of the a-gardener is programmed in the language E of the project Ent [Brom *et al.*, 2005].

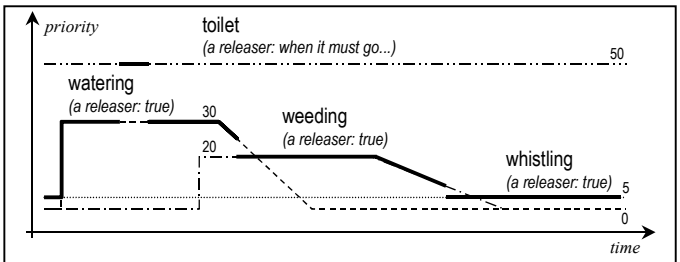


Figure 4: Priorities of top-level goals of the a-gardener. The resulting course of behaviour (without interaction with the user-agent) is indicated by the bold line. The time of switching from *watering* to *toilet* is based on the incremental rate of the internal drive underlying "when it must go" releaser.

5 Results: observed limitations

In this section, we present observed limitations on believability of behaviour of the artificial gardener. The results clearly reveal four main flaws of the a-gardener and thus of the S-HRP method. First, the S-HRP betree does not allow for intentions, the best one could do in S-HRP is to associate intentions with top-level goals. Second, concurrent processes are not allowed—just one a-action can be performed in each time step. Third, some situations require planning, at least to some extent, but S-HRP avoids classical planning. And finally, strong need for transition behaviours, i.e. small processes that applies during task switching, has been recognised. Unfortunately, it is not straightforward to express them in S-HRP. We anticipate that some of these limitations can be avoided by utilizing a different reactive hierarchical method, but others are more fundamental.

5.1 Intentions

Case 1. Choosing an alternative:

N-gardener: Before he starts watering, he decides whether to hose or to water by a can.

A-gardener: When the watering goal becomes active (i.e. *what* to do), the a-gardener is not able to choose between alternatives (i.e. *how* to do), because exactly one process is associated with the top-level goal.

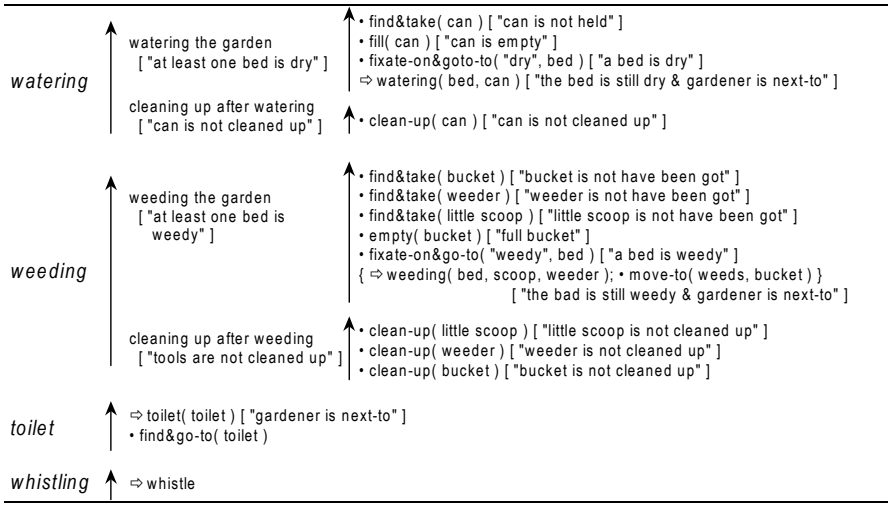


Figure 5: The schema of behaviour of the a-gardener. Parenthesis denotes parameters, brackets denote releasers, angle brackets stand for a sequence. Arrows denote priorities, the p-step with the highest priority is at the top. A-actions are marked with (⇨), subgoals are marked with (•). Contexts are omitted for simplicity.

It is noteworthy, that contrary to other hierarchical reactive methods (e.g. [Bryson, 2001]), the p-steps are written in the reversed-order (the first thing done is on the top) and the releasers are expressed in a negative form. The reason for this is to make the programming simpler.

Case 2. Adding a new intention:

N-gardener: When a neighbour asks for a can, the n-gardener promises her to bring her the can after he finishes watering and then, he becomes *intended* to this task.
A-gardener: It completely lacks capability to add itself a new intention, that means a new top-level goal.

Comment: Changes of intentions and feasibility of choosing alternative ways how to accomplish a goal are basic requirements on action selection model for h-agents. These features are not built-ins of S-HRP, but are well-defined in others hierarchical reactive architectures, namely in the Belief Desire Intention (BDI) [Wooldridge, 2002]. S-HRP could be pushed towards BDI by specifying *simple goal driven hierarchical reactive planning* (S-GHRP):

1. A new structure of *goal* is introduced. It is a quintuple $<\{pr\}, me, r, c, f>$, where $\{pr\}$ is a set of processes that can accomplish the goal, *me* is a procedure for means-ends reasoning among the processes, and *r*, *c* and *f* are a releaser, a context and a priority, a function of time, respectively. A *top-level goal* is just an ordinary goal.
2. An *extended sequence* is a simple sequence of a-actions, processes or goals.
3. A *p-step* is augmented as follows: it is a quadruple $<p, r, c, {}^{g}a>$, where ${}^{g}a$ is an extended action, i.e. a subprocess or an a-action or a goal or an extended sequence.
4. A S-GHRP betree is partially *modifiable on-line*. When an h-agent is running, a new goal can be added to the betree, both top-level one and a subgoal.
5. In S-GHRP, we say that a goal is active, preactive, inactive or sleeping *iff* the p-step encapsulating the goal (or a sequence with the goal) is active, preactive, inactive or sleeping, respectively. Top-level goals are never sleeping, but all top-level goals that are not intended (i.e. not a part of the betree) can be considered as such. From the other hand all not-sleeping goals can be regarded as intentions.
6. The EXEC procedure of the algorithm in Fig. 2 is called also when the *act* variable contains a goal (line (14)) and it performs *me* reasoning in this case (line (25)).

Notice, that the set of all active and preactive elements of the S-GHRP betree is similar to so-called *intention structure* of JAM architecture [Marcus, 1999], which is an implementation of BDI. We suggest that implementations of BDI can be exploited in solving the issue on intentions.

5.2 Concurrent processes and interleaving

Case 3. Pure parallelism:

N-gardener: When he is watering, he whistles from time to time for joy.
A-gardener: It lacks capability to do two tasks simultaneously.

Comment: A simulated body of a believable h-agent should be viewed as a group of semi-independent resources that can perform a-actions concurrently, and S-GHRP should be applied in parallel version, where more active nodes can co-exist in one betree-layer. This idea is hardly surprising and, in fact, a lot of reactive hierarchies address this issue (e.g. [Blumberg, 1996; Bryson, 2001]). It is also noteworthy, that modelling of preferences' combination may be required. That means choosing a compromise action when two (or more) concurrent tasks compete for the same resource (for a discussion on this topic see [Tyrrell, 1993, p. 185-187]).

Case 4. Preparation:

N-gardener: When he is beginning watering, he goes to a chamber and finds a can. Because he knows that he will weed afterwards, he finds also a bucket, and puts a weeder and a little scoop into it. Then he takes the bucket and the can and goes to the garden.
A-gardener: When it is beginning watering, it takes a can from a chamber and goes to the garden. When it finishes the watering, it returns to the chamber for a bucket, a weeder and a little scoop.

Comment: Two goals conflict, active watering and inactive weeding, but even though inactive, weeding has to manifest itself in order to save the second trip to the chamber. Goals have to be *interleaved*. The question is how to give "losers" chances to influence overall behaviour out of their time-slots. Classical solution is to use a planning technique, in this case partial-order simple task network planning (STN) would be appropriate. Unfortunately, it is not straightforward (and perhaps even not biologically plausible) to combine reactive methods with this kind of planning. Therefore, we suggest another solution based on semi-autonomous fuzzy triggers:

1. The priority function of a goal, *f*, is replaced with the set of fuzzy-triggers $\{t^{+}\}$.
2. A *fuzzy trigger* is like a releaser in that it continuously monitors an environment, an agent's body or its mind in order to recognise some relevant situations. Unlike a releaser, the trigger is able to invoke resource negotiation procedure *ne*(*pow*) between an active goal (or goals) and an inactive/preactive applicant. *pow* is the actual power of the trigger (a value <0, 1>).
3. Based on the result of *ne* the applicant can either subsume the active goal, or the active goal can let the "loser" manifest itself shortly, or the *me* procedure of the active goal can switch to another process[1].

The challenging issue is to identify situations that should invoke negotiation. One of such situations is: *an h-agent is attracted by an object that is supposed to be use later*. This notion of semi-autonomous triggers puts S-GHRP a bit towards Minsky's Society of Mind [1985]. A fundamental question on efficiency of this method rises. What is more

[1] We are working on a prototype implementation of negotiation procedure using Soar [Newell, 1990].

computationally effective? STN planning or a bundle of reactive triggers and negotiation procedures?

Case 5. Task inhibition:

(In this case, we assume that a goal of lending a can to the neighbour is intended also by the a-gardener.)
N-gardener: When he finishes watering, he goes to the neighbour and lends her the can.
A-gardener: When it finishes watering, it puts the can to the chamber, then it picks it immediately up and goes to lend it to the neighbour.

Comment: There is another kind of situation that should be recognised by a trigger: *the h-agent attempts to do a task whose effect will be later cancelled*. This situation is unlike Case 4 because the result of negotiation is temporal *inhibition* of a subgoal. To describe this, a new type of trigger is useful: *an inhibition trigger*. It temporarily inhibits a releaser of a p-step that would invoke a conflicting goal (i.e. putting the can to the chamber).

Extended definition of a goal is: the goal is a quintuple $<\{pr\}, me, r, c, \{t^+, t^-\}>$. t^+ and t^- are tuples $<t, ne>$, where t is the trigger and ne is the negotiation procedure. The difference between t^+ and t^- is that t^+ starts negotiation in order to activate the goal, while t^- starts negotiation in order to inhibit a releaser of a p-step with an undesirable goal.

Inhibition is a fundamental feature of architectures like of Maes [1991] (an inhibition link) or Minsky [1985] (a suppressor and a censor agent). Its need is also mention for example in [Charles *et al.*, 2002]. Nevertheless, it is typically not a built-in of reactive hierarchies. We will call this kind of extended S-HRP *negotiatory goal driven hierarchical reactive planning*, the N-GHRP.

5.3 Stop and think

Case 6. Seeing a distance:

(This case extends the scenario from Section 2 as it describes one situation more precisely.)
N-gardener: When he is preparing tools for weeding, he first looks around and then chooses almost optimal way how to pick up a bucket, a weeder and a little scoop.
A-gardener: When it is preparing tools for weeding, it follows the priorities of the p-steps and no matter how the objects are far it always picks up the bucket, then the weeder and finally the little scoop (see Fig. 6).

Comment: This is an observation of a task with complex appetitive behaviour. A question is what we humans do in these situations. We think that two cases should be distinguished. The first is when a human perceive all objects of the interest at once and there are less then three or four of these objects. The second case encompasses more complicated situations with hidden objects or more objects on the scene. We think that in the former case the human *directly* perceive the order of how to pick the objects up[2], while in the latter case the human *consciously* stops and thinks a bit about what to do next.

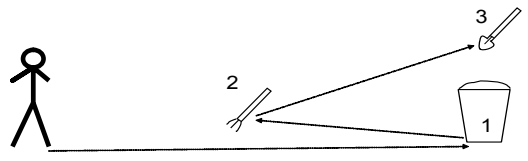


Fig. 6: The order in which the n-gardener picks the objects up is fixed.

For an a-gardener: The latter case calls for a conventional planner that should be invoked by the first subprocess of the process with complex appetitive behaviour. The purpose is to re-arrange the order of appetitive subtasks (i.e. to change priorities of p-steps—it corresponds to stop and think activity). The former, direct perceiving, can be simulated by using releasers and triggers for each possible ordering.

Case 7. A sharp timeout:

(We assume a can is not in the chamber and must be looked for.)
N-gardener: He remembers that the can is often in the chamber. When he does not find it there, he starts searching it within the whole house. As time is passing, he becomes more and more angry. After a while, he wants to give it up, but then he suddenly spot the can in the garage. He picks it up and returns to the garden.
A-gardener: It remembers that the can is often in the chamber. When it does not find it there, it starts searching it within the whole house. After 14.55 minutes of searching, it catches sight of the can in the dining room. It makes two steps towards the can, but just before it reaches the can, the timeout (15 min.) expires. The task of watering is failed.

Comment: A sharp timeout can be expressed directly in S-HRP by a context of a p-step. We suggest that instead of the sharp timeout, a soft one should be used. It incorporates not only time, but also appropriate environmental/body/mental states. N-GHRP triggers serve this purpose better than contexts, because they can facilitate negotiation.

5.4 Transition behaviour

Case 8. No transition:

N-gardener: When nature calls if he is watering, he puts down the can and goes to the toilet.
A-gardener: If it needs to go to the toilet during watering, it goes with the can in its hands and puts it down when in the toilet.

Comment: We have stumbled on so-called *cleaning behaviour*, which is in the case of humans performed in some of its form as a consequent of a consumatory behaviour almost ever. An example of pure cleaning is cleaning up the can. This behaviour can be simply described in S-HRP by adding one p-step after the consumatory act. However, special kind of cleaning behaviours, *transitions*, that mean short behaviours that should automatically apply

[2] Here, we refer to the concept of an *affordance* and *direct perceiving* of James J. Gibson [1979]. However, the discussion on this topic is out of the scope of this paper.

when two "major" ones are being switched, complicate the situation. An example of transition is putting away the can. The need for transitions is noted for example in [Blumberg, 1996; Mateas, 2002]. The problem is that they can not be simply expressed in S-HRP.

We suggest that both for pure cleaning and for transitions negotiation from N-GHRP could be utilised as follows:

1. As the result of negotiation, the incoming goal should give a small amount of time to perform cleaning or transition process to an outgoing goal. The amount of time should be proportional to the ratio of the necessities of behaviours.
2. If an incoming goal is very urgent, transition may be also performed as a part of it. As an example, assume a case of an attack—the gardener might throw the currently holding object at the aggressor, instead of putting it down.

5.5 Other lessons learned

Here, we briefly mention the remnant of observations.

Postponement. When the task *A* is aimed to suspend the task *B* (e.g. eating watering) *postponement* could be negotiated if *B* is almost finished. This is similar to Case 7—when a-gardener is finishing watering, so-called *small variant* of eating could be performed (e.g. eating a tomato from the garden instead of lunching), or watering should not be interrupted at all.

Quantities. Serious problem appears when an h-agent is confronted with huge amount of objects it is potentially interested in. Consider an h-agent aimed to eat a carrot, which needs to be pulled out from a garden bed first—there are hundreds of such carrots in the garden and typical cognitive h-agent's perception system that is designed to perceive all of them, will push this pile into the h-agents' memory. We suggest that in such a situation, the h-agent should instead of perceiving some concrete objects rather see a container, e.g. the garden bed.

Blocking behaviour. The common problem is a situation in which a process *A* shortly corrupts its own context. A correction process *B* (typically a sibling from the betree) could fix the situation, but then *A* corrupts the context again—that only leads to an infinite loop. Consider the a-gardener who must first hold the can to be able to fill it, but in order to turn water on, it must temporarily put it down. It is the same problem as with Herbert, the robot retrieving cans, that blocked by its arm its camera focused on the can, when it had begun to pick the can up [Connell, 1990]. An h-agent must use a memory to remember that it has to avoid execution of the correction process.

6 Discussion: S-HRP vs. related AI models

We have shown several situations in which S-HRP fails as the action selection model for believable h-agents. We conclude that this does not mean the methodology has to be discarded, but rather reviewed instead. Considering the fact that h-agents carry out large number of complex goals in unpredictable and dynamic environments, the hierarchies together with reactive approach must be utilised anyway. There are two main reasons for this. First, hierarchies reduce design complexity. Second, because believable h-agents are aimed for real simulations with several h-agents running on a single PC, their action selection model must be computationally effective (h-agents belong to the field of applied AI, rather then computational psychology or ethology). Therefore, we think that models based on spreading activation in a flat network (e.g. [Maes, 1991]) or in a hierarchical network (e.g. [Tyrrell, 1993; Negatu, 2003]) would not fit, because they generally suffer from combinatorial complexity.

What does it mean to review S-HRP? S-HRP is partially based on Bryson's [2001] basic reactive plans. We suggest that it can be simply extended into S-GHRP by adding the concept of goals. S-GHRP is in fact BDI architecture (e.g. [Wooldridge, 2002]), nevertheless, we suggest that GHRP can be pushed further towards another approaches. Namely, we suggest adding "stop-and-think" planner (but not conventional planning in general) and semi-autonomous triggers that are able to cause resource negotiation, and inhibit an undesirable goal. The second concept is inspired by Minsky's Society of Mind [1986] and Maes [1991]. We have called such architecture "negotiatory goal driven hierarchical reactive planning", the N-GHRP, and we have recommended applying its parallel version. As N-GHRP combines reactive approach with conventional planning, it might be viewed as a *hybrid* architecture representative.

What is the contribution? We see the main contribution of the architecture in that the triggers are able to break the monolithical reasoning procedure into relatively independent modules. Notice, that decomposition of the reasoning procedure is neither decomposition of the body (e.g. [Blumberg, 1996]) nor decomposition of overall behaviour into independent behavioural-modules (e.g. [Bryson, 2003]). It is yet another kind of decomposition.

The decomposition of reasoning blurs the borders between behaviours and makes the alternation among prescripted plans more "smooth" and thus natural and believable (contrary to "rigid 'artificial' switching" in S-HRP and S-GHRP/BDI). For example, sharp timeouts can be avoided, undesirable tasks can be inhibited, preactive behaviour can be demonstrated shortly without its timeslots, transitions can be expressed and postponement can be negotiated.

There is yet another branch of methods that is suitable for believable h-agents. It is any-time planning. For example, Nareyek uses any-time planning based on *structural constrained satisfaction* [2005] and Charles *et al.* exploit a variant of *hierarchical task network planning* and *heuristic search planning* [2002]. We think that anytime planning do not allow for as easy design as reactive hierarchies do. However, the correct comparison between N-GHRP and any-time planning methods is a question for future research.

S-HRP and similar methodologies belongs to the branch of so-called *forward-chaining* methods. To complete the

picture, we must mention Soar architecture [Newell, 1990], which is one of the most known cognitive forward-chaining architectures. Soar is also exploited in h-agents simulations. However, it is rather a powerful programming vehicle, not a design methodology. For example, S-GHRP as well as simple task network planning can be programmed in it.

7 Conclusion

In this paper, we have argued that hierarchical reactive planning is not able to cope with human-like behaviour. We have shown several limitations of this branch of methods through behavioural analysis of an artificial gardener, whose behaviour have been designed according to simple hierarchical reactive planning, the S-HRP.

The main limitations of S-HRP include: 1) impossibility of adding new goals/intentions during execution, 2) the shortage of parallel execution and task interleaving, 3) the impossibility of inhibition an undesirable subtask, 4) fixed-ordered steps in appetitive behaviour, 5) rigid "unnatural" switching between behaviours, which disable for example expressing of transition behaviours and postponement. The first one is the limitation only of the S-HRP method. The second one is the limitation of all the methods that do not allow parallel execution and/or preactive behaviours. The third is the limitation of methods that cannot express inhibition. The last two are limitations of the whole branch of reactive hierarchical family.

We have suggested a solution to overcome these by extending S-HRP to N-GHRP, negotiatory goal driven hierarchical reactive planning. It is a hybrid architecture representative. The precise comparison between N-GHRP and other hybrid approaches, namely any-time planning, is a question for future research.

Acknowledgement

The application Ents was developed as a student project at Faculty of Mathematics—Physics, Charles University, Prague. Thanks to Vladislav Kuboň for supervising the project and to Ondřej Bojar, Milan Hladík, Vojtěch Toman, David Voňka and Mikuláš Vejlupek for their contribution. Thanks also to Rudolf Kryl, Iveta Mrázová, Kamamúra Ryšlink, Terezka Slunečnice, Matěj Hoffmann, Tomáš Bureš and three anonymous referees for their suggestions and comments.

References

[Blumberg, 1996] Bruce M. Blumberg. *Old Tricks, New Dogs: Ethology and Interactive Creatures*. PhD thesis, MIT Media Laboratory, Learning and Common Sense Selection, 1996

[Brom et al., 2005] Ondřej Bojar, Cyril Brom, Milan Hladík, Vojtěch Toman. The Project ENTs: Towards Modeling Human-like Artificial Agents. In *SOFSEM 2005 Communications*, pages 111–122, Liptovský Ján, Slovak Republic, January 2005.

[Bryson, 2000] Joanna Bryson. Hierarchy and Sequence vs. Full Parallelism in Action Selection. In: *The Sixth International Conference on the Simulation of Adaptive Behaviour (SAB00)*, pages 147–156. MIT Press, Ma, Cambridge, USA, 2000.

[Bryson, 2001] Joanna Bryson. *Intelligence by Design: Principles of Modularity and Coordination for Engineering Complex Adaptive Agents.* PhD thesis, Massachusetts Institute of Technology, 2001.

[Bryson, 2003] Joanna Bryson. The Behaviour-Oriented Design of Modular Agent Intelligence. In: *Proceedings of Agent Technologies, Infrastructures, Tools, and Applications for E-Services*, pages 61-79, Springer LNCS 2592, Germany, 2003.

[Charles et al., 2002] Fred Charles, Miguel Lozano, Steven J. Mead, Alicia F. Bisquerra, Marc Cavazza. Planning Formalisms and Authoring in Interactive Storytelling. In: *First International Conference on Technologies for Interactive Digital Storytelling and Entertainment*, Germany, 2002.

[Connell, 1990] Jonathan H. Connell. *Minimalist Mobile Robotics: A Colony-style Architecture for a Mobile Robot*. Academic Press, Cambridge, Ma, 1990.

[Forgy, 1982] Charles L. Forgy. Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem. In: *Artificial Intelligence*, 19, pages 17-37, 1982.

[Ghallab et al., 2004] Malik Ghallab, Dana Nau, Paolo Traverso. Hierarchical Task Network Planning. In: *Automated Planning: Theory and Practice*, chapter 11, Morgan Kaufmann Publishers, San Francisco, Ca, USA, 2004

[Gibson, 1979] James J. Gibson. *The Ecological Approach to Visual Perception*. Boston: Houghton Miffin, 1979.

[Maes, 1991] Pattie Maes. The agent network architecture (ANA). In: *SIGART Bulletin*, 2 (4), pages 115–120, 1991.

[Marcus, 1999] Marcus J. Huber. JAM: A BDI-theoretic mobile agent architecture. In: *Proc. of 3rd International Conference on Autonomous Agents*, pages 236-243, Seatle, USA, 1999.

[Mateas, 2002] Michael Mateas. *Interactive Drama, Art and Artificial Intelligence.* Ph.D. Dissertation. Department of Computer Science, Carnegie Mellon University, 2002.

[Minsky, 1985] Marvin Minsky. *The Society of Mind*. Simon and Schuster Inc., 1985

[Nareyek, 2005] Alexander Nareyek. Project Excalibur. An on-going project on agents for computer games. Homepage: http://www.ai-center.com/projects/ excalibur/

[Negatu, 2003] Aregahegn S. Negatu, Stan Franklin. An Action Selection Mechanism for "Conscious" Software Agents. In: *Cognitive Science Quarterly*, 2, pages 363-386, 2002.

[Newel, 1990] Alan Newell. *Unified Theories of Cognition.* Harward University Press, Cambridge, Masachusetts, 1990.

[Prendinger et al., 2004] Predigner, H., Ishizuka, M. Introducing the cast for social computing: Life-like characters. In: *Life-like Characters. Tools, Affective Functions and Applications*, Cognitive Technologies Series, Springer, Berlin, p. 3-16, 2004.

[Tyrrell, 1993] Toby Tyrrell. *Computational Mechanisms for Action Selection*. Ph.D. Dissertation. Centre for Cognitive Science, University of Edinburgh, 1993.

[Wooldridge, 2002] Michael Wooldridge. *An Introduction to MultiAgent Systems*. John Wiley & Sons, 2002.