Time for AI and Society

Proceedings of the AISB'00 Symposium on How to Design a Functioning Mind

17th-20th April, 2000
University of Birmingham

AISB'00 Convention

17th-20th April 2000

University of Birmingham
England

Proceedings of the AISB'00 Symposium on

How to Design a Functioning Mind

Published by

The Society for the Study of
Artificial Intelligence
and the
Simulation of Behaviour

United Kingdom

http://www.cogs.susx.ac.uk/aisb/

ISBN 1 902956 15 4

Printed at the University of Birmingham, Edgbaston, Birmingham B15 2TT, England.

Contents

Plenary Convention talk relevant to the symposium

Abstracts for Poster Presenters

The AISB'00 Convention

The millennial nature of current year, and the fact that it is also the University of Birmingham's centennial year, made it timely to have the focus of this year's Convention be the question of interactions between AI and society. These interactions include not just the benefits or drawbacks of AI for society at large, but also the less obvious but increasingly examined ways in which consideration of society can contribute to AI. The latter type of contribution is most obviously on the topic of societies of intelligent artificial (and human) agents. But another aspect is the increasing feeling in many quarters that what has traditionally been regarded as cognition of a single agent is in reality partly a social phenomenon or product.

The seven symposia that largely constitute the Convention represent various ways in which society and AI can contribute to or otherwise affect each other. The topics of the symposia are as follows: Starting from Society: The Application of Social Analogies to Computational Systems; AI Planning and Intelligent Agents; Artificial Intelligence in Bioinformatics; How to Design a Functioning Mind; Creative and Cultural Aspects of AI and Cognitive Science; Artificial Intelligence and Legal Reasoning; and Artificial Intelligence, Ethics and (Quasi-)Human Rights. The Proceedings of each symposium is a separate document, published by AISB. Lists of presenters, together with abstracts, can be found at the convention website, at http://www.cs.bham.ac.uk/~mgl/aisb/.

The symposia are complemented by four plenary invited talks from internationally eminent AI researchers: Alan Bundy ("what is a proof?"- on the sociological aspects of the notion of proof); Geoffrey Hinton ("how to train a community of stochastic generative models"); Marvin Minsky ("an architecture for a society of mind"); and Aaron Sloman ("from intelligent organisms to intelligent social systems: how evolution of meta-management supports social/ cultural advances"). The abstracts for these talks can be found at the convention website.

We would like to thank all who have helped us in the organization, development and conduct of the convention, and especially: various officials at the University of Birmingham, for their efficient help with general conference organization; the Birmingham Convention and Visitor Bureau for their ready help with accommodation arrangements, including their provision of special hotel rates for all University of Birmingham events in the current year; Sammy Snow in the School of Computer Science at the university for her secretarial and event-arranging skills; technical staff in the School for help with various arrangements; several research students for their volunteered assistance; the Centre for Educational Technology and Distance Learning at the university for hosting visits by convention delegates; the symposium authors for contributing papers; the Committee of the AISB for their suggestions and guidance; Geraint Wiggins for advice based on and material relating to AISB'99; the invited speakers for the donation of their time and effort; the symposium chairs and programme committees for their hard work and inspirational ideas; the Institue for Electrical Engineers for their sponsorship; and the Engineering and Physical Sciences Research Council for a valuable grant.

John Barnden & Mark Lee

AISB‘2000 Convention: Symposium
How to Design a Functioning Mind
The University of Birmingham
17-18 April 2000

Preface

Background to the symposium

Much research in AI is fragmented: people work on language, or vision, or planning, or learning or mathematical reasoning, without necessarily asking how their models could be combined with others in a fully functioning mind; or they discuss multi-agent systems where the agents have only very simple collections of capabilities.

Much research in psychology is equally fragmented: investigating particular capabilities and how they are affected by environmental factors, or brain damage, or gender, or age, etc.; for instance linguistic or visual or problem solving or memory or motor control capabilities. Moreover such research often produces interesting empirical results without leading to a theory that is deep or precise enough to be the basis for a design for a working system.

Some philosophers also think about these topics and attempt to analyse the concepts involved in talking about minds, or necessary or sufficient conditions for various kinds of mentality, but without doing so at a level that might guide an engineer attempting to design a mind: and some of them produce arguments claiming to show that the task is impossible, but without formulating the arguments in a manner that could convince a computer engineer.

Ethologists study the minds of many kinds of animals and how they differ, but often without asking what sorts of architectural differences might underly the observed differences in behavioural capabilities, social structure, etc.

Biologists and paleontologists study the evolution of systems which include humans and other animals but generally find it much easier to investigate the development of physical form and physical capabilities than the mechanisms of mind.

The purpose of the symposium

The DAM (Designing a Mind) symposium adopts a multi-disciplinary approach to the long term problem of designing a human-like mind, whether for the scientific purpose of understand human minds or some engineering purpose. It is intended to bring together people interested in building bridges between various kinds of partial studies, with the long term goal of understanding, at least in principle, how to build a complete mind.

Researchers in any discipline were invited to submit posters which address these issues, whether in a speculative fashion or by reporting firm results which directly contribute to the long term task. Examples of topics might be proposed include: architectures to accommodate multiple aspects of human mental functioning, or analyses of requirements for such architectures, or a critique of existing architectures on the basis of their functional limitations or inconsistent empirical evidence, or discussions of how important aspects of human minds might have evolved, or analysis of the problems of designing an adult mind vs designing an infant mind which develops into an adult mind, or comparisons between capabilities of different animals which provide evidence for architectural differences, or overviews of major results in neuroscience which have implications for the virtual machine architecture of a mind (e.g. evidence from brain-damaged patients indicating what sorts of separable functional modules exist).

Philosophical papers presenting familiar arguments to prove that the task is impossible were not particularly welcome whereas philosophical arguments which highlight some of the difficulties to be overcome or analyse important conceptual confusions were.

Structure of the symposium

The symposium consists of four main half-day sessions followed by a concluding session. Each of the four main sessions will include presentations of full papers and will end with a discussion period. There will also be sessions for poster presentations. David Lodge is special guest speaker and will give a talk at the end of the afternoon session: "Thinks: a novelist's reflections on the consciousness debate." Marvin Minsky's invited plenary lecture on "Large scale models of mind" on the Monday night is also directly relevant to this symposium.

The final session of the symposium will be a discussion aiming to identify achievements of the symposium and important unsolved problems worth addressing in the near future. It may be useful also to discuss future events of the same kind. Three invited plenary lectures which will be presented after the end of the symposium are also relevant to its aims:

Geoffrey Hinton (Tuesday night) on "How to train a community of stochastic generative models."

Alan Bundy (Wednesday night) on "What is a proof? (The sociological aspects of the notion of proof.)"

Aaron Sloman (Thursday mid-day) on "From intelligent organisms to intelligent social systems: how evolution of meta-management supports social/cultural advances."

This booklet includes the full papers and poster summaries that were received by the deadline for inclusion: not all of them in final form.

Final versions of the papers and poster summaries will be made available at the symposium web site
http://www.cs.bham.ac.uk/research/cogaff/dam00

Organising Committee

Aaron Sloman (programme chair) A.Sloman@cs.bham.ac.uk The University of Birmingham,

John Fox, jf@acl.icnet.uk, Imperial Cancer Research Fund

Brian Logan, bsl@cs.nott.ac.uk, University of Nottingham

Noel Sharkey, n.sharkey@dcs.shef.ac.uk, University of Sheffield

Keith van Rijsbergen, keith@dcs.glasgow.ac.uk, University of Glasgow

Yorick Wilks, y.wilks@dcs.shef.ac.uk, University of Sheffield

Graham Winstanley, G.Winstanley@bton.ac.uk, University of Brighton

The DAM Symposium: How to Design a Functioning Mind
17-18 April 2000

Introduction: Models of Models of Mind

Aaron Sloman
School of Computer Science, The University of Birmingham
http://www.cs.bham.ac.uk/~axs/

Abstract

'Designing a Mind' abbreviated as 'DAM' is easier to type than the full title of the symposium. Many people are working on architectures of various kinds for intelligent agents. However different objectives, presuppositions, techniques and conceptual frameworks (ontologies) are used by different researchers. These differences together with the fact that many of the words and phrases of ordinary language used to refer to mental phenomena are radically ambiguous, or worse, indeterminate in meaning, leads to much argumentation at cross purposes, misunderstanding, re-invention of wheels (round and square) and fragmentation of the research community. It was hoped that this symposium would bring together many different sorts of researchers, along with a well known novelist with ideas about consciousness, who might, together, achieve something that would not happen while they continued their separate ways. This introduction sets out a conceptual framework which it is hoped will help that communication and integration to occur. That includes explaining some of the existing diversity and conceptual confusion and offering some dimensions for comparing architectures.

1 Introduction

It is now common in Artificial Intelligence and Cognitive Science to think of humans and other animals, and also many intelligent robots and software agents, as having an information processing architecture which includes different layers which operate in parallel, and which, in the case of animals, evolved at different stages. This is not a physical architecture, but something more abstract.

In the early days of AI there was far more talk of algorithms and representations than of architectures, but in recent years it has become clear to many people that we also need to understand how to put various parts (including algorithms and representations) together into a larger working system, and for that an architecture is required.

Some computer scientists still use the word 'architecture' only to refer to the physical or digital electronic architecture of a computer, as was common about 20 or 30 years ago, and still is in courses on computer architectures. However the word can also be used to refer to the architecture of a company, a symphony, a compiler, operating system, a theory or a mind. In particular, it can be used to describe any complex system made of coexisting parts which interact causally in order to serve some complex function or produce some behaviour. The parts may themselves have complex architectures. The system and its parts need not be physical. Nowadays the word often refers to non-physical aspects of computing systems, i.e. *virtual machines*. E.g. an operating system or chess program is a virtual machine with an architecture, though it will need to be implemented in a physical system, usually with a very different architecture.

'Information processing' is another term which has both narrow and broad interpretations: some people restrict it to refer to the kinds of bit-manipulations that computers do. However it can be used to refer to a wide range of phenomena in both discrete and continuous virtual machines of various kinds, including acquiring perceptual information about an environment, storing facts, deriving new consequences, searching a memory or database for answers to questions, creating plans or strategies, generating goals, taking decisions, giving instructions or exercising control. As the last two illustrate, not all information is *factual*: there is also *control* information, including very simple on-off control signals, variations in continuous control parameters, labels for actions to perform, and descriptions of what is to be done.

1.1 Information processing models

Thinking of a brain or mind as an information processing system with an architecture is quite old in philosophy, psychology and neuroscience. The early British empiricist philosophers thought of a mind as made of a collection of 'ideas' (experiences) floating around in a sort of spiritual soup and forming attachments to one another. Kant (1781) proposed a richer architecture with powerful innate elements that enable having experiences and learning from from them to get off the ground, along with mathematical reasoning and other capabilities. About a century ago Freud's division of the mind into 'superego', 'ego' and 'id' (among other things) directed attention to a large subconscious component in the architecture, also implicit in Kant's notion of a schema. Somewhat later Craik (1943) put forward the idea that animals build 'models' of reality in order to explore consequences of actions safely without actually performing them (though it is not clear whether he understood the notion of a model in a virtual machine). Popper (e.g. in his 1976 and earlier works) advocated similar mechanisms allowing our mistaken hypotheses to die instead of us.

Recent work has added more detail, some inspired by neuroscience, some by computational models and some by both. Albus (1981, p.184) depicts MacLean's idea of a 'triune' brain with three layers: a reptilian level and two more recently evolved (old and new mammalian) layers. (This may be insulting to intelligent reptiles.) More recently, AI researchers have been exploring a number of variants, of varying sophistication and plausibility, and varying kinds of control relations between layers. For instance, see Nilsson's (1988, Ch 25) account of triple tower and triple layer models, and various models presented at this symposium, including our own distinction between reactive, deliberative and meta-management layers.

It is also now commonplace to construe many biological processes, including biological evolution and development of embryos as involving acquisition and use of information. Perhaps the biosphere is best construed as an information processing virtual machine driven partly by co-evolutionary interactions.

1.2 Prerequisites for progress

Theories about architectures for minds, brains, or AI systems raise a host of problems. One is that superficially similar architectures may have important differences (some described below) that have not been analysed adequately by researchers. As a result there is no systematic overview of the space of interesting or important architectures, or the different types of requirements which architectures may be required to satisfy, against which they can be evaluated. In short there are no adequate surveys of 'design space' and 'niche space' and their relationships. See Sloman (1994, 1998b).

A worse problem is that there is considerable terminological confusion, obscured by the confidence with which people use words and phrases referring to mental states and processes, including, for example, 'belief', 'desire', 'intention', 'consciousness', 'learning', 'emotion', 'personality', 'understanding', and many others.

AI researchers who blithely use mentalistic labels to describe various mechanisms on the basis of shallow analogies were berated long ago by McDermott (1981). However the habit does not die easily.

Moreover, a social psychologist interested in human relations is likely to define 'emotion' so as to cover the phenomena associated with social relationships such as embarrassment, attachments, guilt, pride, loyalty, etc., whereas a brain scientist studying rodents may define the word so that it refers to the brain processes and observable behaviours found in such animals. Other foci of interest lead to yet more definitions of 'emotion' and there are dozens of them in the psychological and philosophical literature. By taking a broader view than any of their proponents, we should be able explain how to accommodate all of these definitions (at least those related to real phenomena) in the same framework in the same general framework.

1.3 Architecture-based concepts

The task of getting a clear overview of the variety of information processing architectures and the problems of clarifying our confused concepts are closely connected.

That is because each architecture supports a collection of capabilities, states and processes, and different clusters of such capabilities and the states and processes define different concepts. For example an operating system that does not support multi-processing cannot support the distinction between thrashing and not thrashing nor does it make sense to ask about its interrupt priority levels. Likewise an architecture for an animal or robot supports a family of mental concepts and different architectures support different families.

Thus we need to be clear about the architectural presuppositions of our concepts. Otherwise, different researchers will focus attention on different aspects of reality, and adopt definitions suited to their interests, not realising that they are ignoring other equally important phenomena, like the proverbial group of blind people each trying to describe an elephant on the basis of what they individually can feel.

It is not hard to convince a blind man that he is in contact with only a small region of a large structure. It is much harder to convince people producing theories of mind that they are attending to a tiny part of a huge system. Psychologists have produced dozens of distinct definitions of 'emotion', and instead of taking this as a clue that there is a range of diverse phenomena which should be given different labels, they often argue about which definition is 'correct'. Our own analysis of various sorts of human emotions has begun to show how in a suitably rich architecture, several different types of processes can occur which correspond to what we sometimes call emotions, which we now distinguish as primary, secondary and tertiary emotions, extending the classification of Damasio and others. See Damasio (1994); Picard (1997); Sloman (1998a, 2000); Sloman and Logan (2000).

2 Deceptive clarity

Evolution has produced brains which, at least in humans, give their owners some information about their own internal processing. This information is deceptively compelling, and often thought to be incapable of being erroneous because it is so direct. We seem to have direct access to our thoughts, decisions, desires, emotions and, above all our own consciousness. This familiarity leads many people to think they know exactly what they are talking about when they engage in debates about the nature of mind, and propose theories about consciousness, experience, awareness, the 'first-person viewpoint', and so on.

However, the diversity of opinions about the nature of the phenomena, especially the widely differing definitions offered by various psychologists, cognitive scientists, brain scientists, AI theorists and philosophers of terms like 'emotion' and 'consciousness', casts serious doubt on the assumption that we all know what we are referring to.

2.1 Two sources of confusion

The confusion has several roots, one of which is the hidden complexity and diversity of the phenomena: the architectural presuppositions of human mentality are extraordinarily complex, and still far from being understood. Moreover there are differences not only between human beings at different stages of development or when suffering from various kinds of damage or disease, but also between humans and different sorts of animals and artefacts. So if mental concepts are inherently architecture-relative the study of mind will require many families of concepts to describe all the phenomena adequately, unlike the study of the physical world. Of course different concepts are required for different levels in the physical ontology, e.g. sub-atomic physics, chemistry, astrophysics, geology, etc. In contrast, concepts of mind involve both differences of levels and differences of architectures at all levels.

Another source of confusion is a common type of philosophical error, namely believing that we have a clear understanding of concepts just because they refer to phenomena that we experience directly. This is as mistaken as thinking we fully understand what simultaneity is simply because we have direct experience of seeing a flash and hearing a bang simultaneously. Einstein taught us otherwise.

From the fact that we can recognise some instances and non-instances of a concept it does not follow that we know what is meant *in general* by saying that something is or is not an instance. There are endless debates about which animals have consciousness, whether machines can be conscious, whether unborn infants have experiences, or whether certain seriously brain-damaged humans still have minds. Our disagreement even over what counts as relevant evidence, is a symptom that our concepts of mentality are far more confused than we realise.

There is no point attempting to resolve such questions by empirical research when we cannot agree on which evidence is relevant. Does wincing behaviour in a foetus prove that it feels pain and is therefore conscious, or is it a mere physiological reaction? How can we decide? Does the presence of a particular type of neural structure prove that the foetus (or some other animal) is conscious, or is the link between physical mechanisms and consciousness too tenuous for any such proof to be possible, as many philosophers have argued?

We can explain why there is so much confusion and disagreement by exposing the hidden complexity of the presuppositions of our ordinary concepts, the diversity of the phenomena referred to, and the indeterminateness of most of our 'cluster' concepts.

2.2 Cluster concepts

Many concepts, besides being architecture-based, are 'cluster concepts', referring to ill-defined clusters of capabilities and features of individuals. If an architecture supports capabilities of types C1, ...Ck and produces processes with features F1, ...Fn, then different combinations of those capabilities and features can define a wide variety of states and processes. But our pre-theoretical cluster concepts lack that kind of precision. For a given mental concept M there may be some combinations of Cs and Fs that definitely imply presence of M, and others which definitely imply absence of M, but there need not be any well-defined boundary between instances of M and non-instances. That is shown by the intense debates about intermediate cases.

This does not mean that there is a fuzzy or probabilistic boundary. Fuzzy boundaries sometimes occur where there is smooth variation and a probabilistic classifier is at work. With cluster concepts there can be clear cases at extremes and total indeterminacy as regards a wide range of intermediate cases, because there has never been any need, nor any basis, for separating out the intermediate cases.

Making all this clear will show how we can define different families of more precise concepts related to the capabilities supported by different architectures. Which definitions are *correct* is a pointless question, like asking whether mathematicians are 'correct' in defining 'ellipse' so as to include circles. Wheel-makers and mathematicians have different concerns.

2.3 Refining and extending concepts

When we have a clear view of the space of architectures that are of interest (including architectures for human-like systems, for other animals, for various kinds of robots and for various sorts of software agents) we can then consider the families of concepts generated by each type of architecture. We can expect some architectures to support some of our mental concepts (in simplified forms) e.g. 'sensing', but not necessarily all of our notions of 'pain', 'emotion', 'intelligence', 'consciousness', etc.

For instance, an insect has some sort of awareness of its environment even if it has nothing like full human consciousness, e.g. if it is not aware that it is aware of its environment. Precisely which sort of awareness

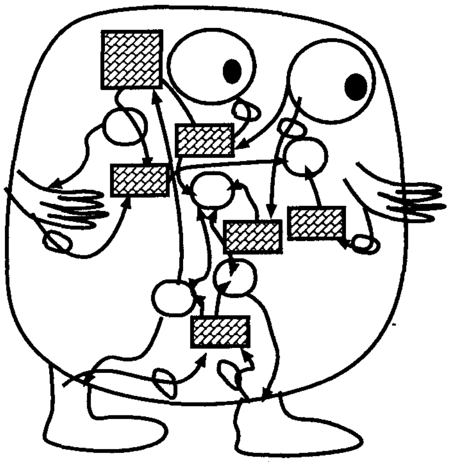


Figure 1: A possible unstructured architecture
In principle, an architecture might be a completely unstructured mess which we could never hope to understand. This is how some view products of evolution. Alternatively evolution, like human designers, may be incapable of producing very complex successful designs unless they have a high degree of structure and modularity, which can provide a principled basis for defining concepts of types of states and processes that can occur.

it has cannot be answered without knowing about its information processing architecture.

Similarly it may be acceptable to use simplified forms of our ordinary concepts in describing some existing AI systems, even though none of them comes close to matching typical human mentality. And if we had a clear idea of the information processing architecture of a foetus at different stages of development, then for each stage we could specify concepts of pain, or awareness that are relevant. However, we should not assume that all concepts applicable to adult humans will be relevant. For instance, it is almost certain that a foetus, or even a new-born infant is not yet capable of being puzzled about the relationship between its mental states and its body or wondering whether a good deity would allow pain to exist. It is possible that a new born infant lacks an architecture capable of supporting wondering about anything.

3 What sorts of architectures?

We know so little about possible information processing mechanisms and architectures (especially the extraordinarily powerful visual mechanisms implemented in animal brains) that it is premature to hope for a complete survey of types of architectures and their capabilities. It could turn out, as some have claimed, that any information-processing architecture produced by millions of years of evolution is bound to be far too messy and unstructured for us to understand as engineers, scientists or philosophers (Figure 1).

Alternatively, it may turn out that evolution, like human designers must use principles of modularity and re-usability in order to achieve a robust and effective collection of architectures, such as we find in many kinds of animals. Figures 2(a) and (b) indicate more structured and modular architectures, combining a three-fold division between perception, central processing, and action, and three levels of processing, with and without a global 'alarm' mechanism. However, such diagrams can be misleading partly because they convey very different designs to different researchers. A frequent confusion is between diagrams indicating state-transitions (flow-charts) and diagrams indicating persisting, interacting components of an architecture. In the former an arrow represents a possible change of state. In the latter it represents flow of information between components. My diagrams are of the latter kind.

To help us understand what to look for in naturally occurring architectures, it may be useful to attempt a preliminary overview of some features of architectures that have already been proposed or implemented. We can then begin to understand the trade-offs between various options and that should help us to understand the evolutionary pressures that shaped our minds.

3.1 Layered architectures

Researchers on architectures often propose a collection of layers. The idea of hierarchic control systems is very old both in connection with analog feedback control and more recently in AI systems. There are many proposals for architectures with three or more layers, including those described by Albus and Nilsson mentioned previously, the subsumption architecture of Brooks (1991), the ideas in Johnson-Laird's discussion (1993) of consciousness as depending on a high level 'operating system', the multi-level architecture proposed for story understanding in Okada and Endo (1992), Minsky's notion of A, B and C brains in section 6.4 of Minsky (1987) and also in several of the papers at this conference.

3.2 Dimensions of architectural variation

On closer inspection, the layering in multi-level architectures means different things to different researchers. There seem to be several orthogonal distinctions at work, which, at present, I can only classify very crudely.
1. Concurrently active vs pipelined layers
In Albus (1981) and some of what Nilsson (1998) writes, the layers have a sequential processing function: sensory information comes in (e.g. on the 'left') via sensors to the bottom layer, gets abstracted as it goes up through higher layers, then near the top some decision is taken, and then control information flows down through the layers and out to the motors (on the other side). I call this an "Omega" architecture because the pattern of information flow is shaped like an Ω. Many AI models have this style. The

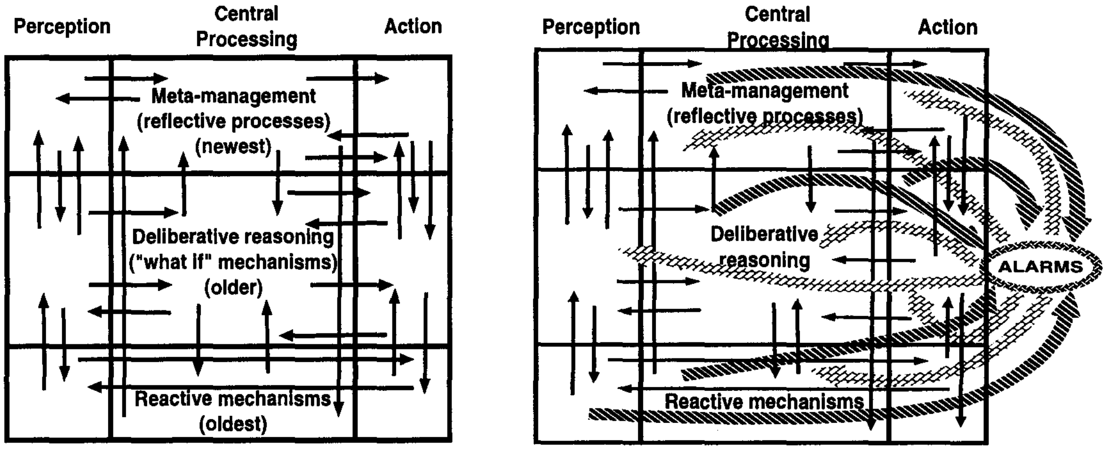


Figure 2: (a) (b)

Nilsson distinguishes 'triple tower' models, with information flowing (mainly) in through a perceptual tower to a central processing tower, then out to a motor tower, and 'triple layer' models where different layers perform different functions. Depending on processing speeds in these mechanisms there may also be a need for a fast global 'alarm' mechanism. Figure (a) serves as a mnemonic indicating the triple tower and triple layer views superimposed, where the various components in the boxes will have functions defined by their relationships with other parts of the system. In (b) a global alarm system is indicated, receiving inputs from all the main components of the system and capable of sending control signals to all the components. Since such alarm systems need to operate quickly when there are impending dangers or short-lived opportunities, they cannot make use of elaborate inferencing mechanisms, and must be pattern based. Global alarm mechanisms are likely therefore to make mistakes at times, though they may be trainable.

enhanced version of Norman and Shallice's "contention scheduling" model, described in Glasspool's contribution to this symposium, is a variant of the Omega schema in which the upward information flow activates a collection of competing schemata where winners are selected by a high level mechanism for controlling attention.

An alternative is an architecture where the different layers are all concurrently active, with various kinds of control and other information constantly flowing within and between them in both directions, as in figure 2 and the 'Cogaff' architecture in 3.

2. Dominance hierarchies vs functional differentiation

A second distinction concerns whether higher levels *dominate* lower levels or merely attempt to control them, not always successfully and sometimes with the direction of control reversed. In the subsumption model (Brooks 1991) higher levels not only deal with more abstract state specifications, goals and strategies, but also completely dominate lower levels. I.e. they can turn lower level behaviour off, speed it up, slow it down, modulate it in other ways, etc. This conforms to the standard idea of hierarchical control in engineering.

By contrast, in a non-subsumptive layered architecture (figures 2 and 3) the 'higher' levels manipulate more sophisticated and abstract information, but do not necessarily dominate the lower levels, although they may sometimes attempt to do so. Higher levels may be able partially to control the lower levels but sometimes they lose control, either via alarm mechanisms or because other influences divert attention, such as sensory input with high salience (loud noises, bright flashes) or newly generated motives with high 'insistence' (e.g. hunger, sitting on a hard chair, etc.). In such a model the *majority* of lower level reactive mechanisms cannot be directly controlled by the deliberative and metamanagement layers, especially those concerned with controlling bodily functions. Some training may be possible, however.

3. Direct control vs trainability

In some layered systems it is assumed that higher levels can directly control lower levels. A separate form of control which is not 'immediate' is re-training. It is clear that in humans higher levels can sometimes retrain lower levels even when they can't directly control them.

For instance, repeated performance of certain sequences of actions carefully controlled by the deliberative layer can cause a reactive layer to develop new chained condition-action behaviour sequences, which can later run without higher level supervision. Fluent readers, skilled athletes, musical sight-readers, all make use of this. (The nature of the boundary between central mechanisms and action control mechanisms is relevant here.)

4. Different kinds of processing vs different control functions

On some models different layers all use the same kinds of processing mechanisms (e.g. reactive behaviours) but perform different functions, e.g. because they operate at different levels of abstraction. In other models there are different kinds of processing as well as different functional roles.

For instance, Figures 2 and 3 present a lowest level that is purely reactive, whereas the second and third levels can do deliberative, 'what if', reasoning, using mechanisms able to represent possible future actions and consequences of actions, categorise them, evaluate them,

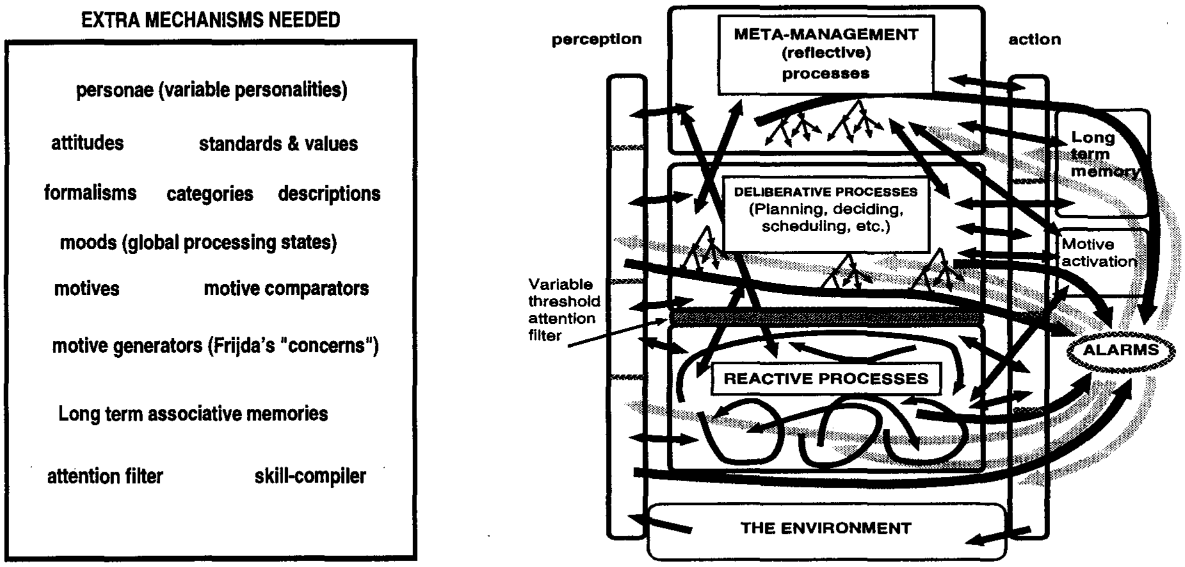

EXTRA MECHANISMS NEEDED
personae (variable personalities)
attitudes
standards & values
formalisms
categories
descriptions
moods (global processing states)
motives
motive comparators
motive generators (Frijda's "concerns")
Long term associative memories
attention filter
skill-compiler
perception
META-MANAGEMENT (reflective) processes
action
Long term memory
DELIBERATIVE PROCESSES (Planning, deciding, scheduling, etc.)
Motive activation
Variable threshold attention filter
ALARMS
REACTIVE PROCESSES
THE ENVIRONMENT

Figure 3: (a) (b)

The Birmingham Cogaff Architecture

We have been exploring ideas based on the collection of mechanisms depicted in Figure 2(b) enhanced with additional components required to make everything work. In (a) we list some additional components required to support processing of motives, 'what if' reasoning capabilities in the deliberative layer, and aspects of self-control. It is conjectured that there is a store of different, culturally influenced, 'personae' which take control of the top layer at different times, e.g. when a person is at home with family, when driving a car, when interacting with subordinates in the office, in the pub with friends, etc. In (b) relations between some of the components are shown along with a global alarm system, receiving inputs from everywhere and sending interrupt and redirection signals everywhere. It also shows a variable-threshold interrupt filter, which partly protects resource-limited deliberative and reflective processes from excessive diversion and redirection. The filter should be thought of as 'wrapped around' the higher levels, with a dynamically varying penetration threshold, dependent, for instance, on the urgency and importance of current tasks.

and make selections. This is not how reactive systems behave. Traditional AI planning systems can do this, and similar mechanisms are able to explain past events, do mathematical reasoning, or do general reasoning about counterfactual conditionals. However, it is possible, indeed likely, that the deliberative mechanisms which go beyond reactive mechanisms in explicitly representing alternative actions prior to selection are themselves *implemented* in reactive mechanisms, which can operate on structures in a temporary workspace.

Reactive mechanisms may be implemented in various kinds of lower level mechanisms, including chemical, neural and symbolic information processing engines, and it is possible that the reliance on these is different at different levels in the architecture. Some kinds of high level global control may use chemical mechanisms which would be too slow and unstructured for intricate problem solving.

Some have argued that human capabilities require quantum mechanisms though I have never seen a convincing account of how they could explain any detailed mental phenomena.

5. Where are springs of action

A fifth distinction concerns whether new 'intrinsic' motives (which are not sub-goals generated in a planning process) all come from a single layer or whether they can originate in any layer. In one variant of the Omega model, information flows up the layers and triggers motivational mechanisms at the top. In other models, processes anywhere in the system may include motive generators, for instance physiological monitors in the reactive layer. The motives they generate may be handled entirely by reactive goal-directed behaviours, or they may need to be transferred to the deliberative layer for evaluation, adoption or rejection, and possibly planning.

6. Handling competing motives

Not all motives will be mutually consistent, so there has to be some way of dealing with conflicts. Architectures differ regarding the locus of such conflict resolution and the mechanisms deployed.

For instance, in some forms of contention-scheduling models, schemata form coalitions and oppositions on the basis of fixed excitatory and inhibitory links in a network, and then some kind of numerical summation leads to selection, which is always done at the same level in the hierarchy. In other models the detection of conflicts might use symbolic reasoning, and the resolution might be done at different levels for different sorts of conflicts.

For instance the decision whether to help granny or go to the marvellous concert might be handled in one part of the system, and the decision whether to continue uttering the current unfinished sentence or to stop and take a breath another way, and the decision to use placatory or abusive vocabulary when addressing some who has angered you might be handled by yet another part of the system.

7. Perceptual to central connections

Architectures with perceptual components differ in the relationships between modes of processing in perceptual modules and more central layers. E.g. is the perceptual processing itself layered, producing different levels of perceptual information to feed into different central layers, or is there a fixed entry level into the central mechanisms, after which the information may or may not be passed up a hierarchy, as in the Omega model. The latter might be described as the 'peephole' model of perception the former the 'multi-window' model of perception.

In 'peephole' perceptual systems, the sensory mechanisms (simple transducers or more complex sensory analysers) produce information about the environment and direct it all to some component of the central architecture. That may trigger processes which affect other parts.

In Figures 2 and 3 it is suggested that the perceptual processes are themselves layered, handling different levels of abstraction concurrently, with a mixture of top-down and bottom up processing, and with different routes into different parts of the central system. For instance deliberative mechanisms may need perceptual information chunked at a fairly high level of abstraction, whereas fine control of movement may require precise and continuously varying input into the reactive system. Differential effects of different kinds of brain damage seem to support the multi-window multi-pathway model, which can also be defended on engineering grounds.

8. Central to motor connections

An analogous distinction concerns the relationship between central and motor processing. Just as there is what I called 'multi-window' perception and 'peephole' perception, so too with action. At one extreme there is only a 'narrow' channel linking the motor system only with the lowest level central mechanism, as in the Omega model: there are motors and they all get signals directly from one part of the central mechanism (analogous to 'peephole' perception). At another extreme there can be a layered, hierarchical motor control system where control information of different sorts comes in directly at different levels, from different layers in the central system.

Humans seem to have motor systems with complex hierarchical skills, and probably also many other animals.

In some proposed architectures (e.g. Albus (1981)) this hierarchical organisation of action is acknowledged, but instead of the action hierarchy being a separate 'tower' (in Nilsson's terminology) communicating with several central processing layers it is folded in to the central control hierarchy. Of course, the two models could describe equivalent systems, but it may sometimes be more useful to think of the central system and the action systems as both having hierarchic organisation. This may help us understand how the whole system evolved in humans and other animals and the increased modularity may help with design tasks. However that is still only a conjecture. Similar comments are applicable to different architectures for perception.

9. Emergence vs 'boxes'

One of the notable features of recent AI literature is the proliferation of architecture diagrams in which there is a special box labelled 'emotions'. Contrast Figures 2 and 3, where there is no specific component whose function is to produce emotions, and instead emotions are explained as emergent properties of interactions between components which are there for other reasons, such as alarm mechanisms and mechanisms for diverting attention (which can happen without any emotion being generated). Elsewhere I have shown how at least three different classes of emotions (primary, secondary and tertiary emotions) emerge in the three layer 'Cogaff' architecture. (This may be compared with the emergence of 'thrashing' in a multi-processing architecture. The thrashing is a result of heavy load and interactions between mechanisms for paging, swapping and allocating resources fairly.)

The problem may be partly terminological: e.g. some theorists write as if all motives are emotions. Then a component that can generate motives may be described as an 'emotion generator' by one person and as a 'motive generator' by another. Separating them accords better with ordinary usage, since it is possible to have motives and desires without being at all emotional, e.g. when hungry. This is just one of many areas where we need far greater conceptual clarity, which may come in part from further study of varieties of architectures their properties, and the state transitions they support.

There are probably many cases whether it is not clear whether some capability needs to be a component of the architecture, or an emergent feature of interactions between components. The attention filter in Figure 3(b) is an example. Instead of a special filtering mechanism, the effects of filtering may be produced by interactions between competing components. The first alternative may be easier to implement and control. The second may be more flexible and general. There are many design tradeoffs still to be analysed.

10. Dependence on language

Some models postulate a close link between high level internal processes and an external language. For instance, it is often suggested (Rolls 1998) that mechanisms analogous to meta-management could not exist without a public language used by social organisms, and in some of Dennett's writings consciousness is explained as a kind of 'talking to oneself'.

A contrary view is that internal mechanisms and formalisms for deliberation and high level self-evaluation and control were necessary pre-cursors to the development of human language as we know it.

The truth is probably somewhere in between, with an interplay between the development of internal facilitating information processing mechanisms and social processes which then influence and enhance those mechanisms, for instance by allowing a culture to affect the development in individuals of categories for internal processes of self-evaluation. (Freud's 'super-ego'). However, it appears from the capabilities of many animals without what we call language, that very rich and complex information processing mechanisms evolved long before external human-like languages, and probably still underpin them. We could extend the word 'language' to refer to forms of internal representation and say that the use of language to think with is prior to its use in external communication.

11. Purely internal vs partly external implementation

A more subtle distinction concerns how far the implementation of an organism or intelligent artefact depends entirely on the internal mechanisms and how far the implementation is shared with the environment. The development in the 70's of 'compliant wrists' for robots, which made it far easier, for example, to program the ability to push a cylinder into a tightly fitting hole, illustrated the advantage in some cases of off-loading information processing into mechanical interactions. Trail-blazing and the design of ergonomically effective tools and furniture are other examples.

From a philosophical viewpoint a more interesting case is the ability to refer to a spatially located individual unambiguously. As explained long ago in Strawson (1959), whatever is *within* an individual cannot *suffice* to determine that some internal representation or thought refers to the Eiffel tower, as opposed to an exactly similar object on a 'twin earth'. Instead the referential capability depends in part on the agent's causal and spatial relationships to the thing referred to. So attempting to implement *all* aspects of mental functioning entirely within a brain or robot is futile: there is always a subtle residue that depends on external relations. (In referring to parts of oneself, or parts of one's own virtual machine the problem is solved internally, as explained in Sloman (1985, 1987).)

12. Self-bootstrapped ontologies

I have been arguing that when we have specified an architecture we shall understand what sorts of processes can occur in it, and will be able to define an appropriate set of concepts for describing its 'mental' states.

However, some learning mechanisms can develop their own ways of clustering phenomena according to what they have been exposed to and various other things, such as rewards and punishments. If a system with the kind of meta-management layer depicted in the Cogaff architecture uses that ability on itself, it may develop a collection of concepts for categorising its own internal states and processes that nobody else can understand because nobody else has been through that particular history of learning processes. The role those concepts play in subsequent internal processing may exacerbate the uniqueness, complexity and idiosyncratic character of its internal processing.

For systems with that degree of sophistication and reflective capability, tscientific understanding of what is going on within it may forevery be limited to very coarse-grained categorisations and generalisations. This could be as true of robots as of humans, or bats Nagel (1981).

4 Human-like architectures

I have tried to bring out some of the design options that need to be faced when trying to explain the architecture of a human mind. When we understand what that architecture is, we can use it to define collections of concepts that will be useful for describing human mental states and processes, though we can expect to do that only to a certain degree of approximation for the reasons in the previous paragraph. However that may suffice to provide useful clarifications of many of our familiar concepts of mind, such as 'desire', 'moods', 'emotion' and 'awareness'.

In particular, so many types of change are possible in such complex system that we can expect to find our ordinary concepts of 'learning' and 'development' drowning in a sea of more precise architecture-based concepts.

We may also be in a better position to understand how, after a certain stage of evolution, the architecture supported new types of interaction and the development of a culture, for instance if the meta-management layer, which monitors, categorises, evaluates and to some extent controls or redirects other parts of the system, absorbs many of its categories and its strategies from the culture. It seems that in humans the meta-management layer is not a fixed system: not only does it develop from very limited capabilities in infancy, but even in a normal adult it is as if there are different personalities "in charge" at different times and in different contexts (e.g. at home with the family, driving a car, in the office, at the pub with mates).

This suggests new ways of studying how a society or culture exerts subtle and powerful influences on individuals through the meta-management processes. The existence of the third layer does not presuppose the existence of an external human language (e.g. chimpanzees may have some reflective capabilities), though it does presuppose the availability of some internal formalism, as do the reactive and deliberative layers.

When an external language develops, *one* of its functions may be to provide the categories and values to be used by individuals in judging their own mental processes (e.g. as selfish, or sinful, or clever, etc.) This would be a powerful form of social control, far more powerful than mechanisms for behavioural imitation, for instance. It might have evolved precisely because it allows what has been learnt by a culture to be transmitted to later generations far more rapidly than if a genome had to be modified. However, even without this social role the third layer would be useful to individuals, and that might have been a requirement for its original emergence in evolution.

We can also hope to clarify more technical concepts. The common reference to "executive function" by psychologists and brain scientists seems to conflate aspects of the deliberative layer and aspects of the meta-management layer. That they are different is shown by the existence of AI systems with sophisticated planning and problem solving and plan-execution capabilities without meta-management (reflective) capabilities. A symptom would be a planner that doesn't notice an obvious type of redundancy in the plan it produces, or subtle looping behaviour.

One consequence of having the third layer is the ability to attend to and reflect on one's own mental states, which could cause intelligent robots to discover qualia, and wonder whether humans have them.

All this should provide much food for thought for AI researchers working on multi agent systems, as well as philosophers, brain scientists, social scientists and biologists studying evolution. I hope the DAM symposium makes some useful contribution to the clarification of these ideas.

Acknowledgements and Notes

This work has benefited from discussions with Brian Logan, Marvin Minsky, and many colleagues and students in the Cognition and Affect project at The University of Birmingham. The work is supported by the Leverhulme Trust. Our papers can be found at

http://www.cs.bham.ac.uk/research/cogaff/

and our tools at

http://www.cs.bham.ac.uk/research/poplog/freepoplog.html

References

James S. Albus. *Brains, Behaviour and Robotics.* Byte Books, McGraw Hill, Peterborough, N.H., 1981.

R. A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

Kenneth Craik. *The Nature of Explanation.* Cambridge University Press, London, New York, 1943.

Antonio R Damasio. *Descartes' Error, Emotion Reason and the Human Brain.* Grosset/Putnam Books, 1994.

P.N. Johnson-Laird. *The Computer and the Mind: An Introduction to Cognitive Science.* Fontana Press, London, 1993. (Second edn.).

Immanuel Kant. *Critique of Pure Reason.* Macmillan, London, 1781. Translated (1929) by Norman Kemp Smith.

D. McDermott. Artificial intelligence meets natural stupidity. In John Haugeland, editor, *Mind Design.* MIT Press, Cambridge, MA, 1981.

M. L. Minsky. *The Society of Mind.* William Heinemann Ltd., London, 1987.

Thomas Nagel. What is it like to be a bat. In D.R. Hofstadter and D.C.Dennett, editors, *The mind's I: Fantasies and Reflections on Self and Soul*, pages 391–403. Penguin Books, 1981.

Nils J. Nilsson. *Artificial Intelligence: A New Synthesis.* Morgan Kaufmann, San Francisco, 1998.

N. Okada and T. Endo. Story generation based on dynamics of the mind. *Computational Intelligence*, 8: 123–160, 1992. 1.

Rosalind Picard. *Affective Computing.* MIT Press, Cambridge, Mass, London, England, 1997.

Karl Popper. *Unended Quest.* Fontana/Collins, Glasgow, 1976.

Edmund T. Rolls. *The Brain and Emotion.* Oxford University Press, Oxford, 1998.

A. Sloman. What enables a machine to understand? In *Proc 9th IJCAI*, pages 995–1001, Los Angeles, 1985.

A. Sloman. Reference without causal links. In J.B.H. du Boulay, D.Hogg, and L.Steels, editors, *Advances in Artificial Intelligence - II*, pages 369–381. North Holland, Dordrecht, 1987.

A. Sloman. Explorations in design space. In A.G. Cohn, editor, *Proceedings 11th European Conference on AI, Amsterdam, August 1994*, pages 578–582, Chichester, 1994. John Wiley.

A. Sloman. Damasio, Descartes, alarms and meta-management. In *Proceedings International Conference on Systems, Man, and Cybernetics (SMC98)*, pages 2652–7. IEEE, 1998a.

A. Sloman. The "semantics" of evolution: Trajectories and trade-offs in design space and niche space. In Helder Coelho, editor, *Progress in Artificial Intelligence, 6th Iberoamerican Conference on AI (IBERAMIA)*, pages 27–38. Springer, Lecture Notes in Artificial Intelligence, Lisbon, October 1998b.

Aaron Sloman. Architectural requirements for human-like agents both natural and artificial. (what sorts of machines can love?). In Kerstin Dautenhahn, editor, *Human Cognition And Social Agent Technology*, Advances in Consciousness Research, pages 163–195. John Benjamins, Amsterdam, 2000.

Aaron Sloman and Brian Logan. Evolvable architectures for human-like minds. In *Proceedings 13th Toyota Conference, on Affective Minds Shizuoka, Japan, Nov-Dec 1999.* Elsevier, 2000.

P. F. Strawson. *Individuals: An essay in descriptive metaphysics.* Methuen, London, 1959.

A Blueprint for a mind by a Categorical Commutative Diagram

Zippora Arzi-Gonczarowski
Typographics, Ltd. Jerusalem, Israel
zippie@actcom.co.il

Abstract

The category of artificial perceptions has been conceived as an infrastructure for the mathematical modelling of AI processes, applying a unified rigorous ontology to various intelligent capabilities. It is shown how this theoretical standard provides an account of perceptual cognitive and affective processes, and integrates them in a high level design proposal for a functioning mind. The schema is presented here through examples, and the formal design is viewed from the perspective of pretheoretical intuitions about minds, cognition, and affect. The dialogue between these intuitions on one hand, and context free mathematical structures on the other hand, is the essence of the research, with the distal goal to approximate a general theoretical account, as well as particular models, of 'minds'.

1 Introduction

A unified ontology that is applied to various intelligent capabilities and skills may help combine them in a fully functioning mind. That ontology should be able to capture some essence of mind processes, yet it should avoid over determinism and be general enough for its eclectic purpose. Mathematical category theory typically provides formal tools to capture a structural essence without being over deterministic. Based on these tools, the category of artificial perceptions has been conceived and proposed as an infrastructure for a theory of AI processes, and it is further proposed to design a high level AI architecture on the basis of the ontology provided by that formalism.

The basic objects of the category are snapshots of perceptual states. Each consists of associating between an environment and some internal structure, producing responses, and recording the experience in a tuned perceptual state, which serves as further basis for more processes, thoughts and deliberation. Streams of perceptual states are formed through transitions that are formalized by morphisms (and other categorical constructs). Any one of the elements that make a perceptual state (i.e. the internal structure, the environment, or the responses) could be modified along paths to other perceptual states. A significant family of transitions involves the formation of complex internal structures, such as acutely perceptive mental representations that could layer on top of basic observations. These complex structures provide a bridge for scaling up to higher-level, rational and emotional, capabilities (e.g. reasoning, creative planning, integrated behaviour management, and autonomous regulatory control). This extended abstract provides a summary of how the schema models the perceptual states themselves, and how it captures various cognitive and affective processes. It is then shown how the unified theoretical standard underlying the various processes enables a rigorous interweaving and integration of all of them in one 'formal mind'. The 'fragments' enhance one another rather than interfere with one another, making a whole that is more than the sum of its parts. The essence of the perceptual-cognitive 'circuits' will be presented by means of examples (the limited length of this paper does not permit a totally self contained digest of all the formal issues, but references are provided to published works).

In mathematical theories, generalizations and principles are typically described by equations. If the concepts and measurement units of several equations match, then they may be embedded in one another, forming an integrated whole. In place of equations, the proposed formalism employs commutative diagrams, that are '*the categorist's way of expressing equations*' (Barr and Wells, 1995, p.83). Like equations, the diagrams can be composed into an integrated compound whole because they are all based on the same categorical premises. The commutative diagram provides a tentative high level 'blueprint'[1] for the eventual programmed design of an artificial 'mind', highlighting the engineering objectives of the formalism. Autonomous action tendencies (urges, emotions) are fomalized as the natural engines of mind vitality: they impel actual performance of transitions between perceptual states. If the diagram provides a 'blueprint of the circuits', then this is the actual 'current'.

The definitions, constructions and results were all operated within the formal mathematical framework, ensuring a tidy treatment that introduces to the related domains tools of mathematical rigor and results that are meticulously stated. On the other hand, the results may be examined relative to the grounding pretheoretical intuitions and existing theories about minds and cognition. After the construction of the diagram, a study of its mathematical

[1] The terminology is borrowed from Magnan and Reyes (1994), who suggest that categorical constructs provide *blueprints* for the design of mind activities.

properties provides further systematizations of intuitions about the boundaries of minds and intelligence.

2 The Working Example

To illustrate the ideas of the proposed formalism, we take off from a perception of a market stand. It consists of a display of fruits, vegetables, flowers, etc. Depending on its sensory motor neural capabilities, an agent may be able to perceive (some of) the colors, tastes, and odors of the produce, feel their touch, listen to the seller. Depending on its current interests and goals the agent may attend to the price of items, to their nutritional value, to their potential use for gastronomic dishes, to how pretty they are going to look on the dinner table, etc. The perceiver may have been trained or programmed to associate each produce with its botanical classification, or with its country of origin. Various perceptual states and attitudes may consist of different apects from the above.

Distinct perceptual states may pertain to different agents, or to distinct states within the same agent. This calls for mind processes that steer between them. A pretheoretical intuition of this study is that adequate steering between perceptual states is an essence of a functioning mind. Mind processes that depend on perception include, among others: urges for direct interaction, that may be either satisfied or not (e.g. get closer, touch, smell, taste), analysis (e.g. "the price of fruit is higher than the price of vegetables"), planning (e.g. plan a nutritious salad, or a novel cross-breeding of fruits.), analogy making (e.g. compare the quality/prices with another stand). Various forms of behaviour (buy, eat, etc.) may follow these processes. Although this toy example is simple, it raises quite a few paradigm issues that a functioning mind needs to tackle. The idea is not original: from the forbidden fruits of the Garden of Eden to sour grapes, our interaction with our natural food has often been a paradigm of other interactions with the world around us.

3 Basic Perceptual Circuits

3.1 Perceptions

The mathematical premises for the proposed formalism have been presented in (Arzi-Gonczarowski and Lehmann, 1998b), and they are briefly summarized now. A *Perception* is defined as a 3-tuple $\mathcal{P} = \langle \mathcal{E}, \mathcal{I}, \varrho \rangle$ where $\mathcal{E}$ and $\mathcal{I}$ are finite, disjoint sets, and ϱ is a 3-valued predicate $\varrho : \mathcal{E} \times \mathcal{I} \rightarrow \{t, f, u\}$.

The set $\mathcal{E}$ represents the perceived environment, *world elements* (*w-elements*) that could perhaps be discerned by a perceiving agent. In the example, each separate produce could be one element of $\mathcal{E}$, or maybe every case of produce would be a w-element, or maybe each stand in the market would be a single w-element: even if the environment exists independent of the perceiving agent, its chunking into individual w-elements varies with perception. To communicate between author and reader, a w-element that stands for a certain object will be designated by w_{bla}, for instance w_{apple}, however the perception under consideration may not necessarily relate to it as an apple.

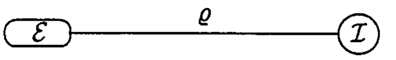


Figure 1: A perception schema

The set $\mathcal{I}$ stands for the discriminations of w-elements, *connotations* that are typically internal to the agent. The market example could feature color connotations (e.g. *red, green* ...), tactile connotations (e.g. *soft, hard* ...), personal taste preferences (e.g. *savoury, unpalatable* ...), nutritional markers (e.g. *vitamin_E, pectin* ...), botanical classifications (e.g. *cucurbitacea, cruciferae* ...), price classes (e.g. *expensive, reasonable* ...) etc. Connotations will be designated in italics. Anything which may be stored and manipulated internally (words, symbols, icons, pictures, diagrams etc.) could be a legitimate connotation.

The 3-valued *Perception Predicate* (*p-predicate*) ϱ relates w-elements with connotations. Let $w_{orange} \in \mathcal{E}$, then a plausible perception $\mathcal{P}$ could have $\varrho(w_{orange}, blue) = f$, $\varrho(w_{orange}, vitamin\_C) = t$, and $\varrho(w_{orange}, Jaffa) = u$ (meaning that, for some reason, perception does not attend to whether this is a Jaffa orange). The p-predicate captures grounded perception. Without going into issues of symbol grounding (Searle, 1984; Harnad, 1990, and others), $\mathcal{P}$ may perhaps associate between lemons and an internal notion of sour taste, namely $\varrho(w_{lemon}, sour) = t$, on the basis of a variety of bottom-up and top-down processes: direct sensation, past experience, the internalization of acquired knowledge, etc.

Actual sets $\mathcal{E}$ and $\mathcal{I}$, and the values of ϱ, once given, provide particular substitution instances. They vary with the embodiment of agents, their environment, their history, their goals, etc.

Hence, the $\mathcal{P}$'s stand for high-level perceptual states that happen at, and above, the level of the basic sensory motor neural apparatus and the recognition of cohesive wholes, where conscious cognizance and symbols begin to play a role. The diagrammatic description of a perception will be based on fig.1: An oval designates a set of w-elements, a circle designates a set of connotations, and the connecting line represents some predicative connection ϱ between the two.

Behaviour at the level of this definition was introduced in (Arzi-Gonczarowski, 1998). It consists of reactions that are motivated by perception, providing for agents that can not only passively perceive, but also respond and interact with the environment. In programming terminology, consider, for example, a w-element $w_{apple} \in \mathcal{E}$, a connotation *juicy* $\in \mathcal{I}$, and assume that $\varrho(w_{apple}, juicy) = t$. The combination of w, α, and $\varrho(w, \alpha)$ could send a message to an object. Methods that can be activated by these messages are the reactions that are associated with perception. A perceived combination of *juicy*, w_{apple}, and

$\varrho(w_{apple}, juicy) = t$, for example, could trigger a gland response and/or a motor reaction: 'bite!'. These responses are part of the definition of a perception $\mathcal{P}$. They may consist of whatever mental or physical actions that the agent is capable of performing.

3.2 Transitions

The flow between perceptions is formalized by *perception morphisms (p-morphisms, arrows)*: Let $\mathcal{P}_1 = \langle \mathcal{E}_1, \mathcal{I}_1, \varrho_1 \rangle$ and $\mathcal{P}_2 = \langle \mathcal{E}_1, \mathcal{I}_2, \varrho_2 \rangle$ be perceptions, then a p-morphism $h : \mathcal{P}_1 \to \mathcal{P}_2$ defines the set mappings: $h : \mathcal{E}_1 \to \mathcal{E}_2$, $h : \mathcal{I}_1 \to \mathcal{I}_2$, and *No–Blur* is the structure preservation condition: for all $w \in \mathcal{E}$ and $\alpha \in \mathcal{I}$, whenever $\varrho_1(w, \alpha) \neq u$ then $\varrho_2(h(w), h(\alpha)) = \varrho_1(w, \alpha)$.

Arrows are a fundamental categorical tool that serves us by capturing a broad spectrum of inter- and intra- agent transitions. The mapping of connotations, $h : \mathcal{I}_1 \to \mathcal{I}_2$, could capture (ı) Simple 'translative' interpretations between perceptions that apply different connotations to the same environment. For example, if $\mathcal{P}_1$ is about pigments, and $\mathcal{P}_2$ is about nutritional substances, then a p-morphism could 'translate' from every pigment to the substance that is most typically associated with it: *h(yellow) = β-carotene*, *h(red) = lycopene*[2], etc. The meaning of 'most typically associated', for that matter, is as captured by the 'no-blur' structure preservation condition. (ıı) Many-to-one mappings could merge similar connotations, for example *h(foliage) = h(leafage) = verdure*, *h(seed) = h(germ) = kernel*, etc. The meaning of 'similar', for that matter, is as captured by the 'no-blur' structure preservation condition. (ııı) The internal representation can be expanded by an inclusive map that is not 'onto'. For example, vitamin connotations could be added to broaden the perspective of perception of produce, capturing the learning of new discriminations. The mapping of w-elements, $h : \mathcal{E}_1 \to \mathcal{E}_2$, could capture (ıııı) Simple, literal, analogies between perceptions that apply the same connotations to distinct environments. If $w_{nuts1} \in \mathcal{E}_1$ stands for all nuts in stand 1, and $w_{nuts2} \in \mathcal{E}_2$ stands for all nuts in stand 2, than a p-morphism could map $h(w_{nuts1}) = w_{nuts2}$, and the same may be done for bananas, etc. (ııııı) A p-morphism could merge w-elements into more general environmental chunks: if A is a subset of indistinguishable w-elements (e.g. all of them are flowers), then a p-morphism could map, for all $w \in A$, $h(w) = w_{flower}$, where w_{flower} is a single, generalized, w-element. (ıııııı) A p-morphism could also expand the perceived environment via an inclusive map that is not 'onto', adding new w-elements to the perception.

The example transitions above are elementary. The idea is that successive basic transitions can be composed into elaborate ones, like a movement of a cartoon character that is made of a series of basic movements of every joint. The 'mind' could perform complex transitions from, say, a perception of the produce in the market to a perception of a nearby exhibit of gems and minerals, creatively soaring from the cherries to a rubi, from the grapes to an emerald, and from the blueberries to a sapphire, making an elaborate interpretive analogy.

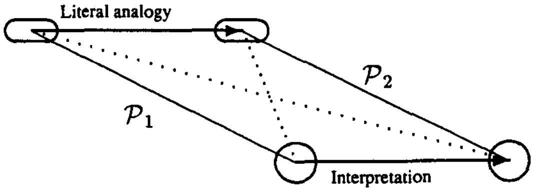


Figure 2: A transition between two perceptions

The diagrammatic description of p-morphism transitions consists of arrows between sets of w-elements and between sets of connotations as in fig.2. Every such transition can be factorized into an *interpretation*, which consists of the mapping of connotations, and a *literal analogy*, which consists of the mapping of environments. They can be composed in any order. That is why they are shown as parallels in the figure. Whether the interpretation (or the literal analogy) is the first or the second factor effects the *metaphorical perception* that is generated in between. The dotted diagonals in fig.2 designate the metaphorical perceptions that blend connotations from one perception with w-elements from another. This was studied in (Arzi-Gonczarowski, 1999b).

Emotive reactions are part of the definition of perceptions, as was just described, hence perceptual states are also affective states. A transition from $\mathcal{P}_1$ to $\mathcal{P}_2$ may involve a change in some, or all, reactions, featuring a change of mood or attitude. If $\varrho_1(w_{grapes}, sweet) = t$, that could trigger the emotive reaction 'take it', but if a change of perception is based on the map *h(sweet)=sour*, and hence $\varrho_2(w_{grapes}, sour) = t$, then the sour grapes would probably conjure a different reaction.

Technically, composition and the identity are defined by those of set mappings, and perceptions with p-morphisms make a mathematical category, designated $\mathcal{P}rc$. This provides a well developed infrastructure for a mathematical theory. The theoretical standard affords constructs that capture perceptual cognitive transitions in a technically rigorous manner. Examples: (ı) 'Blurring' transitions are formalized by traversal of arrows in the reverse direction (in the mirror categroy). This may be applied to cognitive abstraction from details, and to transitions that intentionally 'ignore trifles' that are irrelevant and may interfere with further transitions. For example, the above mentioned transition from w_{cherry} to w_{rubi} needs to first ignore that $\varrho(w_{cherry}, edible) = t$. (ıı) Categorical products and pullback transitions capture joining perceptions into an abstractive schema that highlights the similarities between them and neatly 'blurs' the differences. For example, to cognitively join all the specific stands in the market into an abstractive perception of a 'schematic mar-

[2] Higher level constructs that will be considered later will enable the mapping of a color also to *a disjunction* of substances, if a pigment is associated with more than one substance.

ket stand', one would probably have to ignore the exact layout of displays, that vary from one stand to another. (*iii*) Categorical coproducts (direct sums) capture an expansion of several perceptions into a shared and broader perspective (e.g. a market perception that attends to all the possible connotations offered in section 2). Further pushout transitions formalize 'commonsense' meanings that are shared by all perspectives.

A connecting thread of (*i* – *iii*) above is that perception is fluid and it changes all the time, continuously deleting, replacing, and adding 'facts' and constituents in an ad hoc manner. For example, the edibility of produce could be a crucial discrimination in one context, and a neglectable detail in the transition described above. However, the pretheoretical intuition is that an over permissive account of mind versatililty could deteriorate to inconsistencies that even a 'flexible mind' would have sanctioned. To be implemented in programmed systems, one needs a methodology that is clear, precise, and testable. It should grasp the evasive invariable aspect of meaning with a loose and flexible, yet durable, harness. Structure preservation, in the form of the 'no-blur' condition on p-morphisms, both forces the artificial mind to take a rigorous 'mental note' of the meanings that are being toyed with, and at the same time the formalism also provides flexible tools to neatly play this game.

4 Higher Level Circuits

4.1 Analysis: Representation Generation

Scaling up from basic direct perception to higher-level habilitations is a significant task of the mind. This includes the generation of perceptive and useful representations for reasoning, creative planning, etc. Boolean constructs are proposed for these purposes.

In the example, assume that sets of connotations are closed under Boolean operations. Quite a few features of complemented and distributive lattices, namely Boolean algebras, seem to be capable of serving knowledge representation purposes and related procedural objectives: (*i*) Boolean lattices feature a partial order. This may enable the organization of connotations in *taxonomic hierarchies*, with inheritance of information. For example, *citrus* would probably lie below *vitamin_C*, meaning that 'if it is a citrus then it has vitamin C, and if it does not have vitamin C then it is not a citrus'. Patterns could sometimes be specific to a perception: for example, *orange* could lie below *local*, or *inexpensive*, in one context, but not in another. (A more formal treatment of subsumption of connotations will be provided in section 7.) (*ii*) Boolean lattices feature the two binary operations ∨ and ∧, and the unary operation ¬, allowing the formation of *compound concepts* as Boolean combinations of basic connotations (e.g. *lemon∨orange∨grapefruit=citrus*). (*iii*) The lattice aspect of Boolean algebras provides links for *ease of access*: access the connotations through their links to other connotations (e.g. links from *lemon*, to *citrus*, to *vitamin_C*). (*iiii*) The propositional aspect of Boolean algebras, where ∧ stands for 'and', ∨ stands for 'or', and ¬ stands for 'not' may underlie an intepretation of the representation in logical formulas, and be applied for *ease of inference* (e.g. 'if it is a citrus and it is not a lemon then it must be either an orange or a grapefruit').

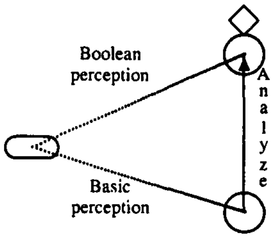


Figure 3: Boolean Representation Generation

Analytic organizations of grounded representations were formalized in (Arzi-Gonczarowski and Lehmann, 1998a) by *Boolean generations*, that close sets of connotations under Boolean operations, transforming the $\mathcal{I}$'s into Boolean algebras (with an adequate embedding of the 3 valued p-predicate in these perceptions). P-morphism are then based on Boolean homomorphisms between connotations, capturing structure aligning transitions. Category theoretical natural transformations systematized the transitions into perceptions that feature the Boolean property. The transition is schematized in fig.3, where the Boolean set of connotations is topped with a diamond. The arrow marked *analyze* designates the natural transformation. Two different canonical Boolean closures will be described in Section 7.

The import of the Boolean construct to behaviour is the option to control and regulate conflicts. A complex combination of perceptual constituents may eventually be wired to a complex combination of conflicting reactions. The lattice structure of Boolean closures provides natural junction collocations for the integration of simultaneous action tendencies. For example, consider a w-element $w_{prickly\,pear}$[3], where $\varrho(w_{prickly\,pear}, sweet) = t$ is likely to conjure the reaction 'eat it', while $\varrho(w_{prickly\,pear}, thorny) = t$ is likely to conjure the reaction 'avoid it'. Perception of the conjunction $\varrho(w_{prickly\,pear}, sweet \wedge thorny) = t$ could be wired to integrative regulatory control of the conflict, such as 'hold the appetitive urge, wear gloves, peel carefully, then eat'. Since the Boolean representation may have a lasting existence inside the agent, integrated reactions can also be planned 'off line', in a deliberative manner (e.g. 'what would I do if I was offered a prickly pear'). This is shared with aspects of design processes that will be discussed in the next subsection.

Reactive control could be wired to anything that the agent is capable of doing, and hence also to *the activation of a p-morphism*, capturing an internal transition to a new affective state to perhaps avoid the conflict by a change of attitude. $\varrho(w_{grapes}, sweet) = t$ may conjure the reaction 'take it', whereas $\varrho(w_{grapes}, too\ high) = t$

[3] A prickly pear, *sabra*, is the edible fruit of certain species of cacti.

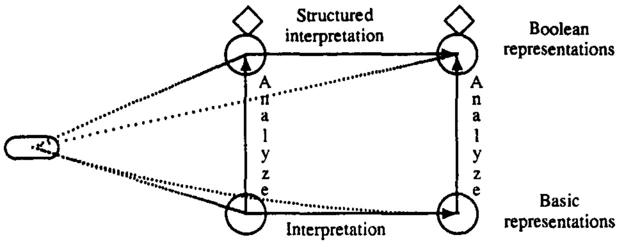


Figure 4: Analytic representations with interpretation

may inhibit that reaction. Perception of the conjunction $\varrho(w_{grapes}, sweet \wedge too\ high) = t$ could be wired to a transition that maps *h(sweet)=sour*, and hence the conflict is eliminated. In the biological context, the need to deal with conflicting action tendencies could have been a significant pressure behind the evolution of an interwined net. It is likely that social agents needed to regulate their behaviour and impulses well before mazy internal connections developed into representations underlying what George Boole entitled 'The Laws of Thought'.

The Boolean construct provides infrastructure, but it does not warrant that integration and control of conflicting action tendencies are always achievable. Possible obstacles: $(\imath)$ Boolean closures have an exponential computational complexity, and minds may be unable to cope computationally with an overwhelming simultaneous combination of too many rousing stimuli. $(\imath\imath)$ Not all conflicts have solutions, and conflicting behaviours can not always be integrated or prioritized. $(\imath\imath\imath)$ Even in view of a plausible compromise or a rational decision, some action tendencies have the nature of not lending themselves to regulatory control, perhaps like the tertiary emotions from (Sloman, 2000). Recall, from Genesis 3:6, the original paradigm of all conflicts: *"And when the woman saw that the tree was good for food, and that it was a delight to the eyes, and a tree to be desired to make one wise, she took of its fruit, and did eat..."*. It may be formalized as $\varrho(w_{fruit}, desired \wedge forbidden) = t$. Minds sometimes need to function in spite of action tendencies that have not been consumed. Until their vigilance is somehow calmed, demanding unsatisfied urges *"crouch at the door"*, and may cause a (partial or total) derailing of mind function, attention and control. In spite of that, it would not be a good idea to do without action tendencies, as they are the essence of vitality, the 'current in the circuits'.

Diagramatically, to scale up the affective-cognitive performance of the agent, the 'plane' that is shown in fig.2 is going to serve as a 'base' for a diagram that looks like a 'box'. The generating arrow of fig.3 is the basic 'corner support'. The functorial construction provides an entire 'wall' along with that 'corner support', in the form of a commutative diagram that is shown in fig.4. The commutative diagram is an *equation* that warrants: A transition from the lower left circle to the upper right diamond can be effected in either one of two *equivalent* ways. In the example: let the lower arrow in fig.4 be a p-morphism h that interprets from a 'vitamin minded' perception $\mathcal{P}_1$, to a 'color' perception of the market $\mathcal{P}_2$, mapping from every vitamin to the color of its reachest source in the stand: $h(vitamin\_A) = red$, $h(vitamin\_C) = yellow$, etc. Let $\mathcal{P}_i{}^\diamond, i = 1, 2$, be the 'diamond' perceptions at the top of the diagram wall, featuring analytic representations that are effected by Boolean closures of the $\mathcal{I}_i$'s as explained before. $\mathcal{P}_1{}^\diamond$ represents, among others, the combined connotation
$multivitamin = vitamin\_A \wedge vitamin\_B \wedge \ldots$, and
$\mathcal{P}_2{}^\diamond$ represents, among others, the combined connotation $colorful = red \wedge yellow \wedge \ldots$. The upper arrow in fig.4 stands for a natural extension of h into a structure preserving interpretation between the two higher level representations, an extension that is part of the functorial categorical construction. It captures a structure aligning transition, where the concept *multivitamin* maps to *colorful*, (indeed, Boolean homomorphisms preserve conjunctions). The arrow path that goes first upwards, and then to the right, stands for a transition that first analyzes and represents 'multivitamin', and then follows with a structured interpretation to 'colorful'. The arrow path that goes first to the right, then upwards, stands for a transition that first follows a simple interpretation from vitamins to colors, and then analyzes and represents 'colorful'. This systematizes the interweaving of analytical and interpretive capabilities in one 'mind', where each capability enhances and supports the other. The result that the two optional transition paths are the same is a way of saying that generations of grounded analytic representations are commensurate, because the schema is methodical.

4.2 Synthesis: Design & Plan Generation

A salient property of the premises is the symmetry between $\mathcal{E}$, the environment, and $\mathcal{I}$, the representation. From a purely technical, context free, point of view, the roles that a w-element and a connotation play in the definitions are interchangeable. This *duality* has the technical consequence that any construct or theorem that is established for connotations (w-elements) can automatically be applied to w-elements (connotations), mutatis mutandis. The duality was applied to erect a second wall that faces the wall from fig.4, formalizing creative–imaginative processes. This was studied in (Arzi-Gonczarowski, 1999a). It is summarized in fig.5, which is dual to fig.4, being technically based on mathematical results that were achieved by sweeping the roles of $\mathcal{E}$ and of $\mathcal{I}$. However, the cognitive processes that are formalized here are different. (The structural similarity between fig.4 and fig.5 is more than a technical convenience. It provides insights into similarities of cognitive processes such as reasoning and design.)

In perceptions with *conceived Boolean environments* the sets of w-elements are Boolean algebras, providing a formalism for an adequate internal conception of combinations of similes and examples from the actual environment. (Boolean environments are designated here by an oval topped with a diamond.) This sets a formal basis for the creative imagination of plans and designs. Transi-

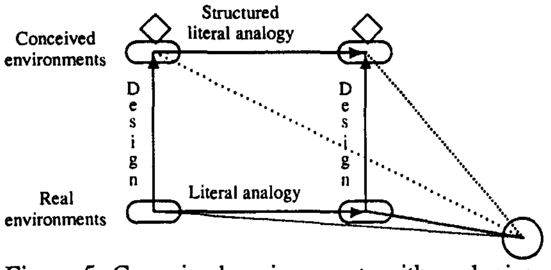


Figure 5: Conceived environments with analogies

tions between perceptions of conceived Boolean environments are based on p-morphisms where the maps between w-elements are Boolean homomorphisms, systematizing structure aligning analogies. Natural transformations formalized methodical cognitive transitions from perceptions of authentic environments to conceived environments. A Boolean combination of w-elements is interpretable as a logical formula that can be further applied for a rigorous effective plan to realize the conceived design.

The functorial construction warrants that, if there exists a simple analogy path between two environments, then this path is preserved by the respective Boolean generations, and can be extended to a Boolean structure preserving path between the conceived environments. This is the import of the diagram in fig.5, that interweaves analogy making and creative design in one 'circuit'. A transition from the lower left oval to the upper right oval (with diamond) can be effected in either one of two ways: One could first conceive of a design and then follow with a structure aligning analogy to another design, or, alternatively, one could first follow a simple analogy between existing environments, and then conceive of a design that is already based on the analogical environment. In the example, assume that the perceiving agent conceives of a basket of mixed fruit. The conceived w-element $w_{cherries} \vee w_{grapes} \vee w_{blueberries}$ is an element in a conceived Boolean environment, and its connotations can be perceived with the 'inner eye', on the basis of what is now offered in the market stand, with authentic w-elements that serve as 'raw material' for the plan. Assume now that there is an analogy from the perception of that authentic environment to another environment, say a literal analogy to another stand in the market. The formalism provides computational tools to extend the bottom arrow to the top arrow between conceived plans, for example a transition to an analogous, planned, basket with fruits bought from the other stand, comparing the overall quality and price.

Action tendencies that are conjured by perceptions of conceived environments systematize 'what if' emotions. An agent that perceives an ulterior environment with its inner eye may have emotive reactions to the possibility that the imagined situation could perhaps become real (e.g. excited anticipation, anxiety). An example could be a motivation to actually effectuate the plan and materialize the design: buy the fruits and arrange them in a real basket. The schema for the generation of conceived environments also systematizes the intuition that both the design and its effectuation are easier if an available basket happens to be perceived in the actual environment.

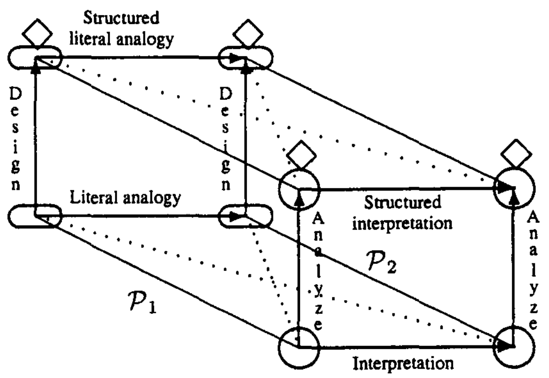


Figure 6: The synthesis

5 The Integrated Circuit

A composite diagram emerges from the fragments: a base with two walls define a box, a whole that features more than the some of its parts. By fig.6, a 'top cover', two 'side walls', and two 'diagonal walls' are gained, representing more perceptions and composite transitions, all of which can be integrated in a single architecture. The category theoretical equational reasoning affirms that the composite box commutes. Various AI cognitive habilitations are interrelated in a wider theoretical framework, with a high-level prescriptive blueprint for an integrated computational framework. Each one of the new walls describes a transition that takes a basic perception ($\mathcal{P}_1$ and $\mathcal{P}_2$, respectively) and scales it up to a cognitive perception with $(\imath)$ Analytic mental representation, $(\imath\imath)$ A perceptive inner eye that conceives of potential designs and plans, $(\imath\imath\imath)$ Integration of behaviours with autonomous regulatory control. The top cover describes an interpretive and analogical transition that applies Boolean homomorphisms to align the high level capabilities $(\imath - \imath\imath\imath))$ that were just described. An example transition of this kind could be based on the interpretive analogy that was mentioned before, between the market display to an exhibit of gems and minerals. Applying the functorial categorical construction, the mind could design, for example, an ornament that would look like a basket of fruits, with rubies for cherries, emeralds for grapes, and sapphires for blueberries. The mathematical construct warrants that one could first conceive of a basket with real fruit, then follow (along the 'top cover' of the box) with a structured interpretive analogy to a conceived ornament or, alternatively, one could first follow a simpler interpretive analogy (along the 'base' of the box) from fruits to gems, and then generate a design on the basis of gems. Further transitions could modify the design by replacement of materials, colors or forms, sometimes to a point where it would not be easy, even to the functioning mind itself, to trace the design back to its original inspiration in the market.

Diagonals and diagonal walls of the diagram have to do with metaphorical perceptions (not all diagonals are shown in the figure). Action tendencies that are conjured by metaphorical perceptions may feature interesting discrepancies: A perceptual state that associates between an environment from one perception and discriminations from another perception, could bring about behaviour that has developed relative to a different, literal, context. It could be quite unexpected in the borrowed context. For example, perceiving an ornament (or reading a research paper) that alludes to fruit may conjure one's appetite, although there is nothing edible there.

6 Inspiriting the circuits

An agent could be initialized to a 'genetic inherited' perceptual state that features essential constituents: it attends to environmental chunks ($\mathcal{E}$) and to discriminations ($\mathcal{I}$) that are vitally consequential to its survival, and its urges and impulses are those that will make it endure. In the biological context these constituents are, of course, naturally selected by evolution. They vary with environments and embodiments. The abundance of natural species, even in specific environments, shows that there is typically more than one rudimentary embodied perceptual state that copes with a situation.

When the initial perceptual state lends itself to contingent transitions, adapts and matures, then perhaps it came with a certain 'mentality'. (Arzi-Gonczarowski, 2000) catalogues the various types of action tendencies that are formalized by the proposed schema, with emphasis on motivations to actually perform the transitions that the blueprint diagram affords. The 'mind' typically functions and develops by interaction with its environment, perceiving and performing the various affective and cognitive transitions that were described above. For example, an agent that perceives how the environment responds to one of its behaviours may be impelled to undergo an internal transition to a modified state that features that behaviour reinforced, or mellowed, according to the perceived response. Different kinds of impact are needed, for instance, for cracking a nut and for peeling a banana. Sensitivity to the properties of materials could be refined through interaction, as well as inter-agent and social sensitivity, and there are, of course, other ripening interactions.

Besides transitions that happen as (either rationally planned, or instinctive) reactions to perceived constituents, some 'mind vitality' could also have its roots in action tendencies that are not related to the agent's relationship with its environment. Perseverant explorer types, for example, are often motivated by persistent drives. A fallout of the formalism is an extension of the spectrum of action tendencies that it systematizes, to behaviour that is driven by internal mental agendas. Internal agendas could be captured as built-in drives towards *attractor states* (although one may never really get to the attractor state)[4]. A formalization of such states is based on 'terminal objects' in mathematical category theory, that will be presented in the next section. Very loosely, an agent with an innate 'curious and interpretive inclination' might have a built-in tendency to move along the arrows of the front wall of the circuit box, invariably analyzing and improving its internal representation. Dually, an agent with an innate 'imaginative designer inclination' might have a built-in tendency to move along the arrows of the back wall of the box, inexorably conceiving and synthesizing novel environments. Subtypes can be formalized by a subtle classifications of arrow routes that are selected.

7 Boundaries of the 'mind'

This section is slightly more technical. It employs basic mathematical tools that are afforded by the formalism to systematize more intuitions about the confines of minds and intelligence. Whether the circuit box is bounded from various directions is the category theoretical version of questions regarding boundary conditions on equations.

7.1 Combinatorial Bounds

In the general case, p-morphisms add new constituents (the exceptions are mergers of similar constituents). Hence, a simple type of bound that may be considered is on the number of different constituents. From the combinatorial point of view, the bound on the number $|\mathcal{I}|$ of connotations are $0 \leq |\mathcal{I}| \leq 2^{|\mathcal{E}|}$ for a given $\mathcal{E}$ (i.e. the possible subsets of w-elements circumscribe the discriminations that one may make). Dually, $0 \leq |\mathcal{E}| \leq 2^{|\mathcal{I}|}$ for a given $\mathcal{I}$ (i.e. the possible subsets of connotations circumscribe the distinct w-elements that one may conceive of). These are obvious bounds along the direction of the arrows.

The category theoretical version of stating that 'one cannot get any further than that' is to show that an object in a category (a perception in $\mathcal{P}rc$) is *terminal.* By definition, a terminal perception T would be such that for all perceptions $\mathcal{P}$, there exists a unique p-morphism $h : \mathcal{P} \to T$. It was shown in (Arzi-Gonczarowski and Lehmann, 1998b) that the *Total Universal Perception of* $\mathcal{E}$, $\mathcal{U}_{\mathcal{E}} = \langle \mathcal{E}, 2^{\mathcal{E}}, \epsilon \rangle$, with $2^{|\mathcal{E}|}$ connotations, has the existence property of (arrows leading to) a terminal object, and this lax[5] terminal object is unique up to isomorphism. This perception has the most evolved representation at the far right of the front wall of the box. Dually, it was shown in (Arzi-Gonczarowski, 1999a) that a similar construct, with the *Universal Environment of* $\mathcal{I}$ that features $2^{|\mathcal{I}|}$ w-elements, has the existence property of (arrows leading to) a terminal object, and this lax terminal object is unique up to isomorphism. This perception has the most evolved conceived environment at the far right of the back wall

[4] A similar idea is offered by the dynamical systems stance in cognitive science.

[5] The uniqueness property of (arrows leading to) this perception does not hold in the 3-valued context.

of the box. These boundary perceptions marry the combinatorial aspect with the categorical algebraic language. They are the *attractor states* from the former section.

The initial object for the category is the *Empty Perception* $\mathcal{P}_\emptyset = \langle \emptyset, \emptyset, \varrho_\emptyset \rangle$. It stands for 'no environment and no representation', and puts a theoretical bound on the 'origin' of the arrows, from the left and from the bottom of the box (perhaps a theoretical *tabula rasa*).

7.2 A Fixed Point Bound

A stronger result for the top cover bound will be shown now, deploying the strengths of the proposed formalism to systematize more intuitions about intelligence. (The introduction of a fixed point formalism in $\mathcal{P}rc$ is new, although it is a direct result of the constructions from (Arzi-Gonczarowski and Lehmann, 1998a).) Figuratively, the top cover of the box could perhaps serve as a base for another box, and the question is whether it is possible to 'pile up' infinitely many boxes, one on top of the other. This would have meant that a mind could infinitely improve its high level capabilities, constantly adding more compound concepts, more plans and designs, and more integrated behaviours.

The vertical arrows of the diagrams are based on perception endofunctors of the form $\mathcal{G} : \mathcal{P}rc \to \mathcal{P}rc$, where $\mathcal{G}(\mathcal{P})$ is a Boolean perception. A vertical arrow $\xi : \mathcal{P} \to \mathcal{G}(\mathcal{P})$ is a natural transformation from the identity functor on $\mathcal{P}rc$ to the functor $\mathcal{G}$. By definition of fixed points for functors[6], a fixed point of $\mathcal{G}$ should be a pair $(\mathcal{P}, h)$ where $\mathcal{P}$ is a perception and $h : \mathcal{G}(\mathcal{P}) \to \mathcal{P}$ is a p-isomorphism. Figuratively, if $(\mathcal{P}, h)$ is a fixed point of $\mathcal{G}$, then $\mathcal{G}(\mathcal{P})$ is the same as $\mathcal{P}$, making 'a wall of no height': the piling up of walls is stopped. This would mean that $(\imath)$ The cognitive transition that is systematized by $\mathcal{G}$ is unable to further scale up perception beyond that which is already featured by $\mathcal{G}(\mathcal{P})$. $(\imath\imath)$ $\mathcal{G}$ is a sensible cognitive process that knows its limitations and is 'aware' of property $(\imath)$.

Two canonical Boolean closures were studied in (Arzi-Gonczarowski and Lehmann, 1998a; Arzi-Gonczarowski, 1999a). Only one features a fixed point. The difference between them is related to validity and completeness in Boolean perceptions. These notions are based on relationships between the Boolean partial order $\leq$ on constituents (connotations, w-elements) on one hand, and perceived lawlike patterns on the other hand. Examples of perceived lawlike patterns could be: 'Inexpensive produce is either seasonal or local', or 'Onions and shallots are the same'. Formally, the perceptual quasi order $\trianglelefteq$ is defined: $(\imath)$ For $\alpha, \beta \in \mathcal{I}$, $\alpha \trianglelefteq \beta$ if $\forall w \in \mathcal{E}\ \ \varrho(w, \alpha) = t \Rightarrow \varrho(w, \beta) = t$ and also $\varrho(w, \beta) = f \Rightarrow \varrho(w, \alpha) = f$. $(\imath\imath)$ For $x, y \in \mathcal{E}$, $x \trianglelefteq y$ is defined in a dual manner. Example lawlike patterns of (Boolean combinations of) constituents:
$\neg \mathit{expensive} \trianglelefteq \mathit{seasonal} \vee \mathit{local}$,
$w_{onion} \trianglelefteq w_{shallot}$ and $w_{shallot} \trianglelefteq w_{onion}$.

As already explained in section 4, since Boolean lattices feature a partial order, this enables the organization of connotations in hierarchies. In a *valid* Boolean perception $\leq\ \subseteq\ \trianglelefteq$, meaning that the formal Boolean hierarchy can be verified by perceptual observations. In a *complete* Boolean perception $\trianglelefteq\ \subseteq\ \leq$, meaning that all observed lawlike patterns are reflected in the Boolean structure. Boolean perceptions are always valid, but not necessarily complete. Perceptions in the *valid and complete* Boolean subcategory, $\mathcal{P}rc^{\text{bl-cmp}}$, feature total internalization of perceived lawlike patterns[7].

The simplest Boolean closure takes the constituents of basic perception as free generators, defines a free functor $\mathcal{G}^{\text{fr}} : \mathcal{P}rc \to \mathcal{P}rc^{\text{bl}}$, and systematizes a general cognitive transition from basic perceptions to Boolean perceptions. It captures methodicalness and open-mindedness, but not perceptual acuity, because $(\imath)$ $\mathcal{G}^{\text{fr}}(\mathcal{P})$ is, in the general case, incomplete (freedom means that there is no dependence between constituents, which is the essence of lawlike patterns). $(\imath\imath)$ $\mathcal{G}^{\text{fr}}$ has no fixed point. In particular, $\mathcal{G}^{\text{fr}}$ is unable to 'sense' a case where $\mathcal{P}$ is already a Boolean perception, and it unconditionally generates a Boolean set of 2^{2^n} constituents over any given n constituents. (A combinatorial explosion will be avoided when the 'pile' eventually hits the general combinatorial upper bound).

The sketch-structure of perceptions (Arzi-Gonczarowski and Lehmann, 1998a) answers the imperviousness of $\mathcal{G}^{\text{fr}}$. Loosely, a p-morphism in the sketch-structured subcategory, $\mathcal{P}rc^{\text{Sk}}$, preserves lawlike patterns, namely the quasi order $\trianglelefteq$ (the technical details can be found in the cited works). The endofunctor $\mathcal{G}^{\text{fr-cmp}} : \mathcal{P}rc^{\text{Sk}} \to \mathcal{P}rc^{\text{bl-cmp}}$ is a free functor. Loosely, it 'moves things around' in the Boolean lattice to reflect the perceived patterns. Consequently, the transition is perceptually acute: $(\imath)$ $\mathcal{G}^{\text{fr-cmp}}(\mathcal{P})$ is valid and complete: it features total observation and internalization of all lawlike patterns that are perceptible by $\mathcal{P}$. $(\imath\imath)$ For all valid and complete Boolean perceptions $\mathcal{P}$, $(\mathcal{P}, \xi^{-1})$ is a fixed point of $\mathcal{G}^{\text{fr-cmp}}$. This is a sensible cognitive process that knows its limitations, it is 'aware' of property $(\imath)$, and would not modify perceptions that it is unable to amend.

The fixed point formalism tells us that $\mathcal{G}^{\text{fr-cmp}}$ is superior to $\mathcal{G}^{\text{fr}}$, not only because it is more perceptually acute, but also because it has an 'awareness' that avoids the 'unnecessary piling up of boxes'. This bound is cognitively derived from within, on the basis of own observations and own intelligence. This is different from the 'bureaucratic' combinatorial bound that has nothing to do with innate perceptual capabilities. Familiar intuitions that have just been systematized are $(\imath)$ Abstract speculations are not enough for real knowledge. A perceptive agent should acutely relate to its environment to construct a truly intelligent knowledge representation. $(\imath\imath)$ Sensible cognitive processes should be aware of their limitations.

Based on the observation that the category of $\mathcal{G}^{\text{fr-cmp}}$–

[6] (See, for example, Barr and Wells, 1995, p.272)).

[7] Detection of lawlike patterns can be based on a programmed implementation like LAD (Boros et al., 1996).

algebras is, in particular, a generalized poset, one gets a hierarchy of valid and complete Boolean perceptions as fixed points $(\mathcal{P}, \xi^{-1})$ of $\mathcal{G}^{\text{fr-cmp}}$. (Figuratively: a sequence of bounded walls of ascending size.) This systematizes the intuition that among perceptions with equally advanced Boolean capabilities (namely $\mathcal{G}^{\text{fr-cmp}}$), those with the more detailed grounding apparatus will generate better cognizance. The initial, empty, perception makes the *least fixed point* (a zero size wall). This captures the intuition that even with the best speculative mind, no true cognition can emerge if there is no grounding apparatus that interacts with an authentic environment. Cognition is both enabled and circumscribed by perception.

Remark: An affective fallout of the acutely perceptive Boolean structures that observe and internalize all law-like patterns, as described above, is a certain gain in introspection. When two observed constituents subsume one another, then they are merged by the structure. For example, if both *half full* $\trianglelefteq$ *half empty* and *half empty* $\trianglelefteq$ *half full*, then a valid and complete Boolean perception would merge the two connotations *half full* and *half empty* into one connotation. Assume now that the generating perception (perhaps that of the owner of the market stand) features a 'positive' emotive reaction when it perceives a half empty case (i.e. half of the merchandise has already been sold), but it features a 'negative' emotive reaction when it perceives a half full case (i.e. half of the merchandise has not been sold). The Boolean representation hence features the cognitive acknowledgement of a self contradicting emotion which is recorded in the representation. The 'mixed feeling' could perhaps be mellowed by a wiring to regulatory control.

8 Summary and Future Work

Lawvere (1994) wrote that categorical constructs approximate a particular model of the general which should be sufficient as a foundation for a general account of all particulars. Hence the schema can be evaluated if it eventually provides infrastructure for the approximation of particular models of minds, on the basis of the general account that has been proposed so far.

The schema still waits to be implemented in a programmed system. Like a reduced instruction set for a computer, the formal ontology conflates the types of building blocks that are required for a high level architecture (w-elements, connotations, p-predicate, categorical primitives, Boolean primitives), but not necessarily the spectrum of mind mechanisms that are modelled.

References

Z. Arzi-Gonczarowski. Wisely non rational – a categorical view of emotional cognitive artificial perceptions. In D. Cañamero, editor, *Papers from the 1998 AAAI Fall Symposium: Emotional and Intelligent*, pages 7–12, Orlando, Florida, October 1998.

Z. Arzi-Gonczarowski. Categorical tools for perceptive design: Formalizing the artificial inner eye. In J.S. Gero and M.L. Maher, editors, *Computational Models of Creative Design IV*, pages 321–354. Key Centre of Design Computing and Cognition, University of Sydney, Australia, 1999a.

Z. Arzi-Gonczarowski. Perceive this as that - analogies, artificial perception, and category theory. *Annals of Mathematics and Artificial Intelligence*, 26(1-4):215–252, 1999b.

Z. Arzi-Gonczarowski. A categorization of autonomous action tendencies: The mathematics of emotions. In *Proceedings of the 15th European Meeting on Cybernetics and Systems Research (EMCSR) Symposium - Autonomy Control: Lessons from the Emotional*, Univ. of Vienna, April 2000.

Z. Arzi-Gonczarowski and D. Lehmann. From environments to representations–a mathematical theory of artificial perceptions. *Artificial Intelligence*, 102(2):187–247, July 1998a.

Z. Arzi-Gonczarowski and D. Lehmann. Introducing the mathematical category of artificial perceptions. *Annals of Mathematics and Artificial Intelligence*, 23(3,4): 267–298, November 1998b.

M. Barr and C. Wells. *Category Theory for Computing Science*. Prentice Hall, 1995.

E. Boros, P.L. Hammer, T. Ibaraki, A. Kogan, E. Mayoraz, and I. Muchnik. An implementation of logical analysis of data. Rutcor Research Report RRR 22-96, Rutgers University, New Brunswick, N.J., July 1996.

S. Harnad. The symbol grounding problem. *Physica*, D 42:335–346, 1990.

F.W. Lawvere. Tools for the advancement of objective logic: Closed categories and toposes. In J. Macnamara and G.E. Reyes, editors, *The Logical Foundations of Cognition*, pages 43–55. Oxford University Press, 1994.

F. Magnan and G.E. Reyes. Category theory as a conceptual tool in the study of cognition. In J. Macnamara and G.E. Reyes, editors, *The Logical Foundations of Cognition*, pages 57–90. Oxford University Press, 1994.

J.R. Searle. *Minds, Brains, and Action*. Harvard University Press, 1984.

A. Sloman. Architectural requirements for human-like agents, both natural and artificial (what sorts of machines can love?). In K. Dautenhahn, editor, *Human Cognition and Social Agent Technology*. John Benjamins Publishing, 2000.

Metaphorically Simulating (Metaphorically Simulating) Self (ABSTRACT)

John Barnden
School of Computer Science
University of Birmingham
England, U.K.
http://www.cs.bham.ac.uk/~jab/

1 Metaphorical Self-Reflection as Option

One aspect of a complete agent, situated within anything like our world, must be the ability to reason about other complete agents and about itself as a complete agent. Now, as cognitive linguists and others have shown, much human discourse concerning agents is highly metaphorical. For example, we commonly talk about each others' minds as being physical containers, of each other as being made up of competing sub-people, of ideas as living creatures, etc. – all in perfectly mundane discourse, not (just) poetry and other literary art. A further tenet is that this discourse rests on metaphorical views that are crucial conceptual processing aids rather than just linguistic icing. Assuming that this is true, people's thinking, not just their discourse, about each other and about themselves is partly, and perhaps highly, metaphorical. It is therefore plausible to suggest that artificial complete agents should reason about each other and themselves in metaphorical terms. Note also that metaphor-based self-reasoning could include metaphor-based self-practical-reasoning, i.e. metaphor-based self-control.

2 Metaphorical Self-Reflection as Practical Necessity?

Indeed, it is even plausible to suggest that there is no practical alternative to doing substantial amounts of reasoning about mental states by means of metaphor, given

(a) the messiness of the world, other agents included,

(b) any complete agent's messiness even as perceived by its own internal reflective processing,

(c) the widely acknowledged ability of metaphor, when used appropriately, to provide more economical and precise description, and more effective reasoning, than is otherwise practical (or perhaps possible) in messy domains.

A further point is that

(d) a system may come gradually to have effective mechanisms of reflection by a process of learning partially based on self-observation; and this process could reasonably be expected to be related to the way we learn how to reason about other agents, a matter that could be partially mediated by existing discourse practices, including the use of metaphor.

3 Metaphorical Self-Reflection at One Remove

Moreover, even if an agent is not inclined to reason about itself on the basis of metaphor, it must have the ability to reason (in whatever way) about people reasoning about it in a metaphorical way. This is self-reasoning at one remove, so to speak.

4 A Relevant Implemented System

Part of the paper will describe a system called *ATT-Meta* that we have developed for conducting metaphor-based reasoning, and that has been applied largely to the special case of metaphor-based reasoning about mental states. For example, it can trace through implications of two ideas being "far apart" in a mind considered as a physical region. The techniques used in the system could also be used reflectively by a complete agent to reason about itself. The system can in addition be applied to reasoning about other agents' metaphorical reasoning, and in particular their metaphorical reasoning about the system itself. The system's metaphorical reasoning is done largely by a procedure akin to the mental-simulation procedure that is popular for reasoning about mental states. Indeed, mental simulation is ATT-Meta's main tool for reasoning non-metaphorically about agents' beliefs and reasoning.

5 Further Observations

Three further, linked, observations raised by the above considerations are as follows.

(1) Dominating Oneself with a Metaphor

If an agent's self-control is partly metaphor-based then, even if a particular metaphorical view that is involved in this self-control inaccurately describes the agent, the self-control may to some extent tend to make the system behave as if it were indeed accurately described by the metaphor. For example, if the metaphorical view fails to be sensitive to particular opportunities for external or internal actions by the agent, then self-control may be deprived of the opportunity for exploiting those possibilities, so that the agent does not in fact perform actions that it could in principle perform. (This means that it would be best to have a plethora of metaphorical views engaged in self-reflection/control, just as there are in ordinary discourse because of the diversity of mental life discoursed about.)

(2) Submitting Oneself to a Metaphor

If a complete agent is self-adaptive in the sense that aspects of it adjust to the activities of other aspects, then we can envisage the possibility of one aspect A of the agent "helpfully" becoming more like the way another aspect B perceives A as being like (much as people act stupidly if you convey that you think they're stupid). Thus, if B perceives A in a (partly) metaphor-based way, then A may adapt to become more like what the metaphorical view would expect. Thus, to continue the broad theme of (1), metaphor-based self-reflection could be to some extent a self-fulfilling prophecy.

(3) Metaphorical Qualia

Consider consciousness, in the maximally meaty sense of felt self-consciousness. If that self-consciousness in an agent partly rests on metaphorical views of itself, including its own mind, then, I claim, the self-consciousness could involve the agent feeling that its own mind is indeed as dictated by the metaphor. For instance, if one of the metaphorical views is MIND AS PHYSICAL CONTAINER, then (to some limited extent, and not all the time) the agent's mind could feel to the agent like a physical container. This feeling would be part of the reality of consciousness (because any feeling is automatically part of that reality), even though the metaphor would undoubtedly be misdescribing the underlying nature of the agent.

References

http://www.cs.bham.ac.uk/ jab/ATT-Meta/

Making Modularity Work: Combining Memory Systems and Intelligent Processes in a Dialog Agent

Joanna Bryson
MIT AI Lab; Cambridge MA 02139; USA
joanna@ai.mit.edu

Abstract

One of the greatest obstacles to designing a mind is the complexity of integrating different process types, time frames and representational structures. This paper describes a methodology for addressing this obstacle, Behavior Oriented Designed (BOD), and explains it in the context of creating an agent capable of natural language dialogue.

1 Introduction

Modularity of some degree and description is almost universally accepted in modern understandings of the operation of human minds. On a physical level, we know a great deal about the different structures and functions of various elements of the central nervous system. The spinal cord, the neocortex, the hippocampus, the amygdala, the cerebellum, the lateral geniculate nucleus, the various sense organs — while our understanding of these systems is not complete, we have begun to know their individual architectures and their contributions to intelligence as a whole. Similarly, we are developing a set of fairly well described psychological modules we know to be at least partly independent — declarative knowledge, motor skills, episodic memory, drives, emotions, perception, recognition.

Controversies surrounding modularity focus not so much on its existence as on its nature and extent. The question of modularity is not whether it exists, but how it is organized. For example, are modules necessarily fully encapsulated (that is, disconnected from each other) as in Fodor (1983) and Brooks (1991), or can they draw information from each other, as indicated by Karmiloff-Smith (1992) or Ramachandran and Blakeslee (1998). In AI, there is a further question — how can we as creators of intelligence create this order?

This paper describes a methodology for addressing this problem, Behavior Oriented Designed (BOD), and explains it in the context of creating an agent capable of natural language dialogue. The goal is to be able to create a system capable of perception and action; of maintaining both local behavioral coherence and global dedication to multiple, possibly conflicting goals; of learning; and, in this case, of expressing itself. This requires the combination of parallelism with ordered sequential behavior, and modularity with coherence. It requires the integration of both memory and intentionality across several different time frames.

This paper begins with a full description of BOD. We then illustrate the methodology by describing the construction of a dialog agent. We have been pursuing research in this area for two very different, but not necessarily exclusive purposes. The first is to leverage the current state-of-the-art in the design of complex agents to the problem of creating tutorial dialogs. The second is to incorporate semantic and syntactic information gathered through statistical natural language processing into an intentional dialogue agent. Both of these projects are work in progress, but both effectively illustrate important aspects the problem of designing a mind. The paper concludes with a brief discussion of what additional features might be required in BOD for constructing a true "mind," and whether BOD is compatible with these extensions.

2 Behavior Oriented Design (BOD)

2.1 Behaviors and Behavior-Based Design

Behavior Oriented Design is a methodology for constructing complex agents. It is designed to be applicable under any number of languages and most popular agent architectures. As can be gathered from its name, BOD is a derivative of Behavior-Based Artificial Intelligence (BBAI) (Brooks, 1991; Maes, 1991; Matarić, 1997), informed by Object Oriented Design (OOD) (e.g. Coad et al., 1997). Behavior-based AI is an approach that specifies that intelligence should be decomposed along the lines of perception and action. Behaviors are described in terms of sets of actions and the sensory capabilities necessary to inform them. This sensing must inform both *when* the actions should be expressed, and *how*. In other words, there are really two forms of sensing: sensing for detecting context, and sensing for parameters and feedback of motor actions.

The central observation of behavior oriented design is that mere sensing is seldom sufficient for either detecting context or controlling action. Rather, both of these abilities require full perception, which in turn requires memory. Perception exploits experience and expectation to perform discriminations more reliably than would otherwise be possible. This observation has two consequences in the BOD methodology. First, memory becomes an essential part of a behavior. In fact, memory requirements serve as the primary cue for *behavior decomposition*, — the process of determining how to divide intelligence into a set of modules. This strategy is analogous to the central tenet of object-oriented design, that process is best described and ordered in terms of state.

Although behaviors should be autonomous in so far as they provide for their own awareness of appropriateness to context, they should not necessarily be so sufficiently informed as to know whether their current operation is in line with the intentions or behavioral context of the entire agent. This is the other consequence of acknowledging the role of expectation in perception. The intentions of the agent are themselves state, and as such are the domain of a separate module dedicated to arbitration between behaviors. This process is known as action selection.

2.2 Action Selection in BOD

Action selection, unless sufficiently informed, turns into a combinatorially explosive search process, such as productive planning (Chapman, 1987). Behavior oriented design addresses this problem in two ways. First, in common with standard BBAI, the behaviors themselves control many of the details of action, thus significantly reducing the potential search space. Second, BOD relies on reactive planning. Reactive planning provides the expertise of experience and expectation in the form of preprogrammed plan elements, which are executed as seems appropriate based on the agent's perceptions. These perceptions are again based in the behaviors. As this indicates, BOD does not take the strictly encapsulated view of modularity, but rather allows for a well-defined interface between behavioral modules. The form of this interface is also taken from OOD. The interface is built of methods on the objects that represent the agent's behaviors.

Reactive planning provides for the sequencing of behavior through the appropriate execution of reactive plans. Reactive plans are not themselves limited to sequential structure, but generally also exploit hierarchy, with a number of different possible plan elements ready for execution at any particular time. Behavior oriented design provides for both of these means of ordering behavior. *Action patterns* are simple sequences of action and sensing primitives. *Competences* are prioritized collections of plan elements, the operation of which will tend to achieve a particular goal. Most plan elements will also carry perceptually-dependent preconditions for determining whether the element can and should operate. When a competence is active, its highest priority element that can operate is activated.

Consider an example in blocks world. Let's assume that the world consists of stacks of colored blocks, and that we want to enable an agent to meet the goal of holding a blue block. The perceptual operations in this plan are based on the visual routine theory of Ullman (1984), as implemented by Horswill (1995). A possible plan would be:

1	(holding block) (block blue)	*goal*
2	(holding block)	drop-held-object
3	(fixated-on blue)	grasp-top-of-stack
4	(blue-in-scene)	fixate-blue

In this plan, the highest priority plan element is at the top: recognizing that the goal has been achieved. The competence will terminate if either the goal is achieved or if no elements can execute. Otherwise, the highest priority element is executed. Consider the case where the world happens to consist of a stack with a red block sitting on the blue block. If the agent has not already fixated on the blue block before this competence is activated, then the first operation to be performed would be element **4**. Otherwise, if for example the previously active competence has already fixated on blue, **4** would be skipped. Once a fixation is established, element **3** will trigger. If the grasp is successful, this will be followed by element **2**, otherwise **3** will be repeated. Assuming that the red block is eventually grasped and discarded, the next successful operation of element **3** will result in the blue block being held, at which point element **1** should recognize that the goal has been achieved, and terminate the competence.

Infinite retries can be prevented through a number of means: either through habituation at the element level, timeouts at the competence level, or through a separate attentional mechanism which is triggered by repeated attempts or absence of change. As this last potential mechanism indicates, BOD's action selection mechanism also provides for the switching of attention between multiple possible drives. Indeed, a parallel mechanism is necessary in order for an agent to be sufficiently reactive to operate in a dynamic world (c.f. Georgeff and Lansky, 1987; Matarić, 1997; Bryson, in press). In BOD's action selection, this mechanism takes the form of a special root competence that keeps track of the basic drives for the agent and monitors them for changes in focus of attention. I refer to systems with these characteristics as having Parallel-rooted, Ordered Slip-stack Hierarchical (POSH) action selection. These attributes have been used in a stand-alone architecture, Edmund (Bryson and McGonigle, 1998; Bryson, 1999b) and incorporated into other architectures: Ymir, a multi-modal character architecture (Thórisson, 1999; Bryson, 1999a; Thórisson and Bryson, in preperation) and PRS, a leading reactive architecture for agents and robots (Georgeff and Lansky, 1987; Bryson, in preperation).

2.3 The Design Process

The analogy between BOD and OOD is not limited to the obvious implied metaphor of the behavior and the object, as discussed in Section 2.1, nor to the use of methods on the behavior objects for specifying the interface to the reactive plans, as explained in Section 2.2. The most critical aspect of BOD is its emphasis on the design process itself. As in OOD, BOD emphasizes cyclic design with rapid prototyping. The process of developing an agent alternates between developing libraries of behaviors, and developing reactive plans to control the expression of those behaviors. As in OOD, BOD provides guidelines not only for making the initial behavior decomposition, but also for recognizing when a decomposition has turned out to be inadequate, and heuristic rules for correcting these problems.

2.3.1 The Initial Decomposition

The initial decomposition is a set of steps. Executing them correctly is not critical, since the main development strategy includes correcting assumptions from this stage of the process. Nevertheless, good work at this stage greatly facilitates the rest of the process.

The steps of initial decomposition are the following:

1. Specify at a high level what the agent is intended to do.
2. Describe likely activities in terms of sequences of actions. These sequences are the the basis of the initial reactive plans.
3. Identify an initial list of sensory and action primitives from the previous list of actions.
4. Identify the state necessary to enable the described primitives and drives. Cluster related state elements and their primitives into specifications for behaviors. This is the basis of the behavior library.
5. Identify and prioritize goals or drives that the agent may need to attend to. This describes the initial roots for the POSH action selection hierarchy.
6. Select a first behavior to implement.

The lists compiled during this process should be made either in a notebook, or better in computer files, which can serve as documentation of the agent. If the documentation is kept in files, the use of a revision control system is strongly recommended, in order to record the project's history.

Experience has shown that documentation is most likely to be maintained if it is in fact a functional part of the code. For this reason, code files for the separate elements of the system should be kept segregated by function. In particular, most reactive architectures require special definitions of the coded primitives. These definitions should be kept in a single dedicated file, and the primitives sorted by the behavior module in which they are implemented. This file being functional is therefore always up-to-date, and a clear and convenient documentation of the interface.

In selecting the first behavior, it is often a good idea to choose a simple, low-level priority that can be continuously active, so that the agent doesn't "die" immediately. For example, on a mobile robot with a speech synthesizer, the bottom-most priority of the main drive hierarchy might be a "sleep" function, which keeps track of the time and snores every 30 seconds or so. This way, the developer has a clear indication that the robot's control has not crashed, but that none of its interesting behaviors can currently trigger.

2.3.2 The Development Process

The remainder of the development process is not linear. It consists of the following elements, applied repeatedly as appropriate:

- coding behaviors,
- coding reactive plans,
- testing and debugging code, and
- revising the specifications made in the initial phase.

Usually only one behavior will be actively developed at a time. Again, using revision control is a good idea, particularly if multiple developers are working on behaviors.

Reactive plans grow in complexity over the development time of an agent. Also, multiple reactive plans might be developed for a single platform, each creating agents with different overall behavior characteristics, such as goals or personality. It is best to keep all working plans in a special library, each commented with the date of its development, a description of its behavior, and a record of any other plan or plans from which it was derived. A library of historic plans can be used as a testing suite if any radical change is made to the behavior library.

Testing should be done as frequently as possible. Developing in languages that do not require recompiling, such as perl and lisp, significantly speeds the development process, though it may slow program execution time.

2.3.3 Revising the Specifications

The most interesting part of the BOD methodology is the set of rules for revising the specifications. As in OOD, one of the main goals of BOD is to reduce redundancy. If a particular plan or behavior can be reused, it should be. If only part of a plan or an action primitive can be used, then a change in decomposition is called for. In the case of the action primitive, the primitive should be decomposed into two or more primitives, and the original action replaced by a plan element. Ideally, the new plan element will have the same name and functionality as the original action. This allows established plans to continue operating without change.

In general, the main design principle of BOD is *when in doubt, favor simplicity.* A primitive is preferred to an action sequence, a sequence to a competence. Heuristics like the above can then indicate when the simple element must be broken into a more complex one.

If a sequence sometimes needs to contain a cycle, or often does not need some of its elements to fire, then it is really a competence, not an action pattern. A competence may be thought of, and designed, as a sequence of behaviors that might need to be executed in a worst-case scenario. The ultimate (last / goal) step is the highest priority element of the competence, the penultimate the second highest and so on. Triggers on each element determine whether that element actually needs to fire at this instant. If a competence is actually deterministic, if it nearly always actually executes a fixed path through its elements, then it should be simplified into a sequence.

Competences are really the basic level of operation for reactive plans, and a great deal of time may be spent programming them. One way a competence can flag a need for redesigning the specification is by relying on large numbers of triggers. Perception should be handled at the behavior level; it should be a skill. A large number of triggers should be converted into a single perceptual primitive. Another problem can be that too many elements are added into the competence. This makes design more difficult by increasing the probability that a design fault might result in plan elements that operate against each other, unsetting each other's preconditions. More than seven elements in a competence, or difficulty in appropriately prioritizing or setting triggers, indicates that a plan needs to be decomposed into two plans. If several of the elements can be seen as working to complete a subgoal, they may be moved into another competence which replaces them as an element of the parent plan. If two or more of the elements always follow each other in sequence, they should be removed and made into an action pattern, which is again substituted into the original competence. If the competence is actually trying to achieve its goal by two different means, then it should be broken into two sibling competences which are both inserted into the competence's parent plan, with appropriate triggers to determine which one should operate.

2.4 Differences from Related Approaches

Given that this section has emphasized the analogies between BOD and other related approaches, it may be useful to also quickly outline some relevant differences. It should be noted first that BOD and its competitors are methodologies, not just algorithms. In most cases, it should be at least *possible* to solve problems under any approach. The difference is how easy (and consequently, how likely) it is to solve problems using a particular strategy.

There are two main benefits of BOD over standard BBAI: BOD's use of hierarchical reactive plans, and BOD's methodology of behavior decomposition.

Having explicit reactive plans built as part of the architecture greatly simplifies control. When one particular set of behaviors is active (say a robot is trying to pick up a teacup) there is no need to worry about the interactions of other unrelated behaviors. The robot will not decide to sit down, or relieve itself, or go see a movie unless it is at a reasonable juncture with the tea cup. On the other hand, it may drop the cup if something truly important happens, for example if it must fend off an attack from a large dog trying to knock it over. It is much easier to express this information in a reactive plan than to build complex mutual inhibition systems for each new behavior every time a behavior, as is necessary in conventional BBAI. In mutual inhibition or reinforcement systems, the control problem scales polynomially, with explicit plans the problem scales linearly.

What BOD offers in terms of behavior decomposition over other BBAI methods is:

- A better place to start. Instead of trying to determine what the units of behavior are, the developer determines what information the agent is going to need. This is one of the chief insights from OOD.
- A better way to fix things. Unlike other BBAI approaches, BOD does not necessarily assume that decomposition is done correctly on the first attempt. It provides for cyclic development and neat interfaces between behaviors and control.

Although BOD is based on OOD, it is not a fully object-oriented approach. OOD tends to be useful for passive reactive systems, but is used less frequently for designing systems that are actively internally motivated. The addition BOD provides over OOD is the reactive plan component. This allows the expression of motivation and priority as part of the organization of behavior. BOD applies techniques for building plans and decomposing behaviors that are analogous to, but not exactly the same as the OOD methodologies for designing object hierarchies. In BOD the behaviors are not hierarchical, the reactive plans are.

Another recently developed methodology is Agent Oriented Design (Iglesias et al., 1999). AOD assumes that every module is an agent, with intentions and fully encapsulated state. This differs from BOD, which allows different behaviors to have access to each other's state, and models intentions and planning on a global level, across behaviors, not within them. AOD is actually more analogous to standard BBAI than to BOD.

3 Behavior Decomposition and Plan Construction for Dialogue

Behavior oriented design was developed as a methodology to address the scaling issue in behavior based artificial intelligence. So far it has been applied to problems in blocks world, mobile robotics (Bryson and McGonigle, 1998), artificial life (an animal in a simulated ecosystem) (Bryson, 1999b), and virtual reality character development (Bryson, 1999a). However, none of these applications illustrate BOD at a level of cognitive complexity that suits the conventional implications of "mind". For this reason, this paper focuses its illustrations on work currently in progress in the application domain of natural language dialogue.

Dialog systems currently require an enormous amount of engineering, and typically result in relatively brittle systems. We are currently exploring the use of reactive planning in general and BOD in particular for simplifying dialogue system design.

3.1 Selecting Initial Primitives: the First Plan

We begin by considering as an example the problem of dialog management in a system such as TRAINS-93 (Allen et al., 1995). This system was a major effort in addressing the complete problem of dialog, including having a system capable of planning and acting as well as discussing its plans and acquiring its goals verbally. The TRAINS system served as an assistant to a manager attempting to make deliveries of commodities, such as bananas and orange juice, to a number of different cities. In addition, various cities had various important resources, such as trains, cars, processing plants and raw commodities. These cities were connected by rail, so transport requires scheduling in both time and space.

To build a dialog system similar to TRAINS-93, we first list a rough set of capabilities we expect the agent will have. In this case, we can use the existing system as a guide, and assume that the agent will eventually need the same set of speech acts as capabilities. While we are organizing the gross behavior of the agent, these speech acts will be simple primitives that merely indicate their place in execution by typing their name. This practice of implementing bare, representative functionality as a part of early design is called *stubbing* in OOD. Based roughly on TRAINS speech acts, the initial list of primitives is the following:

accept or **reject** a proposal by the dialog partner,
suggest a proposal (e.g. a particular engine or location for
a particular task),
request information (e.g. a particular of the current plan),
supply-info in response to a request, and
check for agreement on a particular, often necessary
due to misunderstandings.

```
(if my-turn)
   (if request-obligation) (if check-request false) reject
   (if request-obligation) (if check-request true) accept
   (if inform-obligation) supply-info
   (if comprehension-failure) check last-utterance
   (if bound-non-requirement)
      (if requirement-checked) check-task
      check-requirement
   (if requirement-not-bound)
      pick-unbound-req , suggest-req
   (if (no task)) request-task
wait
```

Table 1: In this table, indentation indicates depth in the plan hierarchy. Notice that the action primitives generally assume deictic reference, where the perception primitive has set attention to a particular task or requirement.

Working from these primitives, we can construct a high-level plan for dialog management in just a few lines (see Table 1). Here, sensory checks for context are indicated by parenthesis. The primitive actions listed above are in bold face.

The highest level concern for this plan is simply whether the agent should take a turn, or whether it should wait quietly. Once it has decided to take a turn, the highest priority behavior is to fulfill any discourse obligations, including the obligation to try to understand the previous statement if it was not successfully parsed. If there are no existing obligations, the next highest priority is to resolve any inconsistencies in the agent's current understanding, indicated here by having a requirement not entailed by the task bound to some value. This indicates a need either for clarification of the requirement, or of the current task.

If there are no such inconsistencies, but there is an outstanding task to perform, then the next highest priority is to complete the task, which in the case of TRAINS usually involves assigning a particular resource to a particular slot in the problem space. Finally, if there is no task, then this agent, having no other social or personal goals, will seek to establish a new one.

This simple plan indicates a number of elements of state the agent is required to keep track of. These elements in turn indicate behaviors the agent needs to have established. To begin with, the agent needs to know whether it currently believes it has the turn for speaking. Although that may be a simple of bit of information, it is dependent on a number of perceptual issues, such as whether the dialogue partner is actively speaking, and whether the agent itself has recently completed an utterance, in which case it might expect the other agent to take some time in processing its information. The agent may also be capable of being instructed to wait quietly. Further, that waiting might also be time bounded.

3.2 Building a Behavior Library and Drive Structure

To a first approximation, the primitives used in the plan above can be arranged into behaviors as shown in Figure 1.

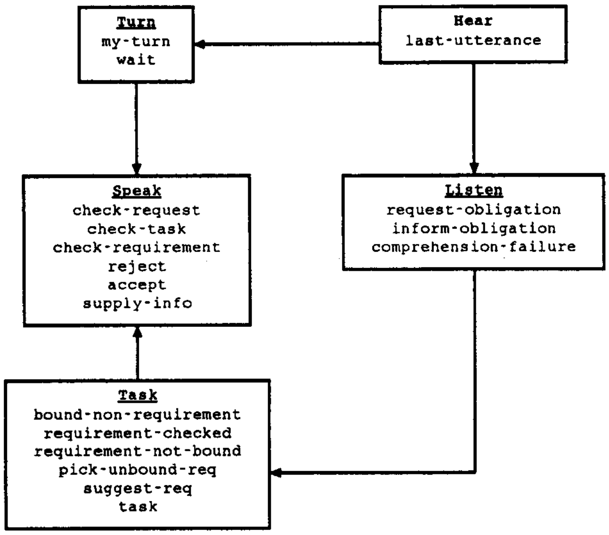


Figure 1: A first cut at a behavior decomposition for a TRAINS-93 type dialog agent. Each box represents a behavior and the primitive senses and actions associated with it. Arrows indicate information flow between behaviors.

The constructive planning required by TRAINS-93 can also be replaced by a fairly short reactive plan (omitted for space) though still supplemented by an $A*$ search algorithm for finding the nearest resources. This suggests that a reasonable initial drive structure for the TRAINS-like dialog agent might be:

1	(if noise)	listen
2	(if need-answer)	think
3	(if my-turn)	take-turn
4		wait

This small plan serves as the parallel operating root of the action selection for the entire dialog agent. The plan in Table 1 fits under the label *take-turn* while the reactive plan for scheduling (including the call to the search algorithm) fits under *think*. A drive structure like this allows another speaker to interrupt, since *listen* has the highest priority. The entire system still relies on the basic behaviors shown in Figure 1. The act of attempting to take a turn would set the flag for "need-answer" if a problem requiring domain-specific planning has been encountered. Solving such a problem should unset the flag, so that turn taking might again operate. Notice that the drive structure has no goal, so will never terminate due to success. Also, the lowest priority element has no precondition, so the drive might never terminate with failure, unless *wait* has a timer and a limit after which *wait* itself fails.

3.3 Scaling the System

The above system obviously hides a great deal of complexity: the problems of parsing the dialog input and constructing sensible output are completely untouched. On the other hand, a BOD system is sufficiently modular that these procedures may be primitives or "black boxes," since many AI systems for language parsing and generation have already been constructed.

Our intention is to use behavior oriented design to organize dialog management for an even more complex system than the one shown above. Our problem domain is tutoring basic electricity and electronics, and we hope to integrate systems that are capable of a wide range of behaviors for assisting students. Examples of desired behavior include analyzing incorrect answers in order to diagnose the learning failure, and providing multi-turn, Socratic method tutoring to lead the students to correcting their basic misconceptions. To be useful with real students, this system will need to be sufficiently reactive to allow the student to either solve the problem prematurely, and also be able to branch into a greater depth of explanation in response to a query or further errors from the student. The design specifications of this tutoring system are described further in (Core et al., 2000).

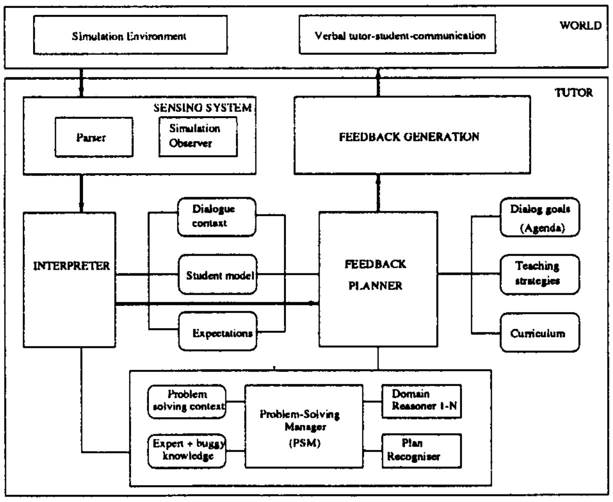


Figure 2: An example of an architecture for a dialogue-based intelligent tutoring system. After Core et al. (2000).

4 Discussion: Can BOD Build a Mind?

The previous sections have presented a methodology for creating complex agents, and an example blueprint for an agent capable of the particularly humanoid capability of natural language dialogue. But can behavior oriented design be used to create a *mind*? This section addresses this question in three different ways. First, we examine the biological plausibility of systems such as those created under BOD. Next, several important elements of mind not explicitly addressed in the previous example are discussed. Finally, we discuss learning.

4.1 Biological Plausibility

BOD hypothesizes the following:

1. most of intelligence is broadly modular,
2. arbitrating between modules requires a specialized mechanism for action selection,
3. complex behavior requires hierarchical and sequential structure for action selection, and
4. switching attention from complex behavior to new salient features or events also requires a specialized mechanism, operating in parallel.

None of these hypotheses have been completely established in the biological literature, however all of them currently have active support in the neuroscience literature. Modularity was discussed in the introduction to this paper. It is also supported by brain cell recording experiments showing that individual cells are associated with different stimuli and/or behavior, and indeed are members of different ensembles, depending on the animal's current context (e.g. Skaggs and McNaughton, 1998). Redgrave et al. (in press) have put forward the hypothesis that the basal ganglia is the specialized organ for action selection in vertebrates. The amygdala has long been implicated as a brain organ dedicated to detecting emotionally salient features in the environment and gating behavior in response to them (Carlson, 1994). Of the hypotheses, the most contentious is probably the third, which I have discussed at length elsewhere (Bryson, 2000). There is considerable evidence that at least some species-typical behavior sequencing is stored in various areas of the vertebrate midbrain (Carlson, 1994).

Another interesting biological analog to the mental architecture constructed under BOD is Rensink's recent theory of vision and visual attention (Rensink, 2000). Rensink proposes that the visual scene is essentially covered with *proto-objects* which are monitored in parallel by the vision system, while only one item is fully attended to at any given time. That item is constructed of approximately four "fingers" of attention which bind proto-objects into the attended, fully represented object. Only attended objects can appear in episodic memory, or be associated with time, though proto-objects may communicate location and gist, particularly on the level of priming. Rensink's work is an interesting parallel, particularly in that it focuses on perception, while BOD focuses on action, and the two models are more or less reciprocal.

4.2 Additional Humanoid Subsystems

As indicated towards the end of the dialog example above, any number of modular systems might be incorporated into a BOD system. This paper and BOD in general emphasize the organization of action and memory. However, a fully humanoid mind might require a number of other systems.

4.2.1 Spatial Action and Perception

An embodied agent, whether embodied physically or in virtual reality, encounters significant challenges not present in a text-based world. For example, a robot typically has noisy and unreliable sensors which require memory and sensor fusion to disambiguate perception. Of course, natural language can also be ambiguous and require multiple knowledge sources. As it happens, BOD's behavior library structure and heuristic programming methodology were originally developed around the problems of mobile robot sensing, and have proven successful in that domain.

Coherent action, particularly of effectors with many degrees of freedom, is easier to perform with an extension to BOD. This extension is a separate intelligent scheduler which selects appropriate motions once the main planner has selected the target positions or gestures. This module is at least roughly functionally equivalent to the cerebellum in vertebrates, in that it handles smoothing of behavior without handling its planning. BOD has been successfully combined with another architecture which has this scheduling feature, Ymir (Thórisson, 1999), although unfortunately very little work has been done so far with this hybrid architecture.

4.2.2 Emotions

Although BOD provides explicitly for motivation and drive, the above system has no specific provision for emotion or affect. In vertebrates, emotions serve as specialized mechanisms for focusing attention, including by deactivating large sections of the cortex (Damasio, 1999). They can also provide complex reinforcement signals for learning behavior (Gadanho, 1999). These functional considerations can be addressed from within BOD. However, a specialized system for mimicking more exactly human or animal emotions would clearly be useful for certain kinds of social interactions — both for making the agent more comprehensible to humans, and for allowing the agent to better model (and therefore perceive) human behavior.

4.2.3 Consciousness and Explicit Knowledge

The dialog system above makes no explicit distinction between conscious and unconscious behavior. Norman and Shallice (1986) propose that consciousness is a sort of special attention which, when activated by some form of interrupt or exception-handling system, aids with a particularly difficult task. This sort of system might be modeled in BOD, perhaps with the more generally associated links to highly plastic episodic memory serving as the specialized attentional systems. BOD systems are capable both of focusing attention on important tasks, and of storing and manipulating records of episodes. However, currently no deliberate model of consciousness has been built under BOD. In fact, the action selection system that operates aspects of the dialog system that are usually unconscious in humans is identical to that which operates elements that are usually perceived as being conscious: the only difference is which behaviors' elements are being manipulated.

4.3 Learning

However powerful a design methodology is, it would always be useful to automate at least some part of the design process by using machine learning. BOD's architecture provides bias useful for enabling certain kinds of learning. Of course, bias and constraint are equivalent: at least one class of problem is very difficult to address within the constraints of the BOD architecture. Nevertheless, BOD can serve as a good starting place for designing minds, including designing minds that learn.

4.3.1 Learning Within Behaviors

BOD is deliberately designed to enable learning within a behavior: in fact, the rate at which state varies is one of the chief cues for what state should be clustered into a particular behavior. Similarly, machine learning techniques can be used for constructing a behavior, or at least part of its state. We are currently engaged in attempting to incorporate statistically acquired semantic lexicons (e.g Lowe, 1997) into a dialogue agent. This could quickly broaden the scope of the agent's ability to recognize conversational contexts. An agent with this lexicon could recognize entire classes of semantically similar sentences for any one programmed interaction.

Similarly, we would like to incorporate the statistically acquired mechanisms of natural language generation of Knight (Knight and Hatzivassilogon, 1995; Oberlander and Brew, in press) into our dialog agent. This would allow us to vary the generative output of our system to be appropriate for various audiences simply by training the mechanism on an appropriate corpus.

Ultimately, it would be interesting to attempt to learn dialog patterns directly from corpora as well. In this case, we could create a "learning Eliza" with only basic turn-taking mechanisms built into the system. The system might be vacuous, but this might not be apparent in gossip-level conversations. We believe, for example, that it might be possible to simulate an individual suffering from William's syndrome this way.

4.3.2 Learning New Plans

Facilitating the learning of new reactive plans was another of the design intentions of BOD. However, this intended feature has not yet been exploited in practice. Learning action patterns should be possible, provided their representation is inserted into a behavior rather than left in a privileged location of the architecture. Learning could occur by trial and error, with generation of new trials done by mutation or recombination of existing patterns. New patterns could also be provided socially, either by imitation or instruction, or by informed search as in conventional planning.

4.3.3 Learning New Behaviors

Although learning new behaviors is clearly a human capacity, this ability is the one most firmly outside of the BOD specification. Allowing for behavioral change would require a complete re-representation of behaviors into some form of generic object class, or better a generic, fine-grained substrate. However, learning acquired modularity in neural representations is currently only in its infancy. There will probably be a significant interval before such techniques can model the complexity of behavior shown in BOD systems. Even if these techniques are finally developed, BOD or a similar architecture might be used to develop the initial agent with its first set of behaviors and competences.

Acknowledgments

Thanks to Jered Floyd, Lynn Stein, Johanna Moore, Claus Zinn, Peter Wiemar-Hastings, Massimo Poesio, Will Lowe and particularly Mark Core.

References

James F. Allen, Lenhart K. Schubert, George Ferguson, Peter Heeman, Chung Hee Hwang, Tsuneaki Kato, Marc Light, Nathaniel Martin, Bradford Miller, Massimo Poesio, and David R. Traum. The TRAINS project: a case study in building a conversational planning agent. *Journal of Experimental and Theoretical Artificial Intelligence*, 7(1):7–48, 1995.

Rodney A. Brooks. Intelligence without reason. In *Proceedings of the 1991 International Joint Conference on Artificial Intelligence*, pages 569–595, August 1991.

Joanna Bryson. Creativity by design: A behaviour-based approach to creating creative play. In Frank Nack, editor, *AISB'99 Symposium on Creativity in Entertainment and Visual Art*, pages 9–16, Sussex, 1999a.

Joanna Bryson. Hierarchy and sequence vs. full parallelism in action selection. In Daniel Ballin, editor, *Intelligent Virtual Agents 2*, September 1999b.

Joanna Bryson. The study of sequential and hierarchical organisation of behaviour via artificial mechanisms of action selection. MPhil Thesis, University of Edinburgh, 2000.

Joanna Bryson. Fundamentals of reactive planning. in preperation.

Joanna Bryson. Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, in press.

Joanna Bryson and Brendan McGonigle. Agent architecture as object oriented design. In Munindar P. Singh, Anand S. Rao, and Michael J. Wooldridge, editors, *The Fourth International Workshop on Agent Theories, Architectures, and Languages (ATAL97)*. Springer-Verlag, 1998.

Niel R. Carlson. *Physiology of Behavior*. Allyn and Bacon, Boston, 5 edition, 1994.

David Chapman. Planning for conjunctive goals. *Artificial Intelligence*, 32:333–378, 1987.

Peter Coad, David North, and Mark Mayfield. *Object Models: Strategies, Patterns and Applications*. Prentice Hall, 2nd edition, 1997.

M. G. Core, J. D. Moore, C. Zinn, and P. Wiemer-Hastings. Modeling human teaching tactics in a computer tutor. In *Proceedings of the ITS'00 Workshop on Modelling Human Teaching Tactics and Strategies*, Montreal, 2000. under consideration.

Antonio R. Damasio. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt, 1999.

Jerry A. Fodor. *The Modularity of Mind*. Bradford Books. MIT Press, Cambridge, MA, 1983.

Sandra Clara Gadanho. *Reinforcement Learning in Autonomous Robots: An Empirical Investigation of the Role of Emotions*. PhD thesis, University of Edinburgh, 1999.

M. P. Georgeff and A. L. Lansky. Reactive reasoning and planning. In *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI-87)*, pages 677–682, Seattle, WA, 1987.

Ian D. Horswill. Visual routines and visual search. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, August 1995.

C. A. Iglesias, M. Garijo, and J.C. Gonzalez. A survey of agent-oriented methodologies. In J.P. Müller, M.P. Singh, and A.S. Rao, editors, *The Fifth International Workshop on Agent Theories, Architectures, and Languages (ATAL98)*, pages 185–198, Paris, 1999. Springer Verlag.

Annette Karmiloff-Smith. *Beyond Modularity: A Developmental Perspective on Cognitive Change*. MIT Press, Cambridge, MA, 1992.

K. Knight and V. Hatzivassilogon. Two-level, many-paths generation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguists (ACL95)*, pages 252–260, 1995.

Will Lowe. Meaning and the mental lexicon. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, Nagoya, August 1997. Morgan Kaufmann.

Pattie Maes. The agent network architecture (ana). *SIGART Bulletin*, 2(4):115–120, 1991.

Maja J. Matarić. Behavior-based control: Examples from navigation, learning, and group behavior. *Journal of Experimental & Theoretical Artificial Intelligence*, 9 (2/3):323–336, 1997.

Donald. A. Norman and Tim Shallice. Attention to action: Willed and automatic control of behavior. In R.Davidson, G. Schwartz, and D. Shapiro, editors, *Consciousness and Self Regulation: Advances in Research and Theory*, volume 4, pages 1–18. Plenum, New York, 1986.

Jon Oberlander and Chris Brew. Stochastic text generatin. *Philosophical Transactions of the Royal Society of London, Series A*, 358, in press.

V. S. Ramachandran and S. Blakeslee. *Phantoms in the brain: Human nature and the architecture of the mind*. Fourth Estate, London, 1998.

P. Redgrave, T. J. Prescott, and K. Gurney. The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, in press.

Ronald A. Rensink. The dynamic representation of scenes. *Visual Cognition*, 7:17–42, 2000.

W.E. Skaggs and B.L. McNaughton. Spatial firing properties of hippocampal ca1 populations in an environment containing two visually identical regions. *Journal of Neuroscience*, 18(20):8455–8466, 1998.

Kristinn R. Thórisson. A mind model for multimodal communicative creatures & humanoids. *International Journal of Applied Artificial Intelligence*, 1999.

Kristinn R. Thórisson and Joanna Bryson. Flexible behavior-based planning for multimodal characters capable of task-oriented dialogue and action. in preperation.

Shimon Ullman. Visual routines. *Cognition*, 18:97–159, 1984.

A Narrative Architecture for Functioning Minds: A Social Constructionist Approach

William Clocksin
Computer Laboratory, University of Cambridge
wfc@CL.cam.ac.uk

Abstract

We aim to build a conceptual framework for artificial intelligence (AI) that gives priority to social relationships as a key component of intelligent behaviour. It starts from Weizenbaum's premise that intelligence manifests itself only relative to specific social and cultural contexts. This is in contrast to one prevailing view, which sees intelligence as an abstract capability of the individual based on a mechanism for rational thought. The new approach is not based on the idea that the mind is a rational processor of symbolic information, nor does it require the idea that thought is a kind of abstract problem-solving with a semantics that can be understood independently of its embodiment. Instead, priority is given to affective and mimetic responses that serve to engage the whole organism in the life of the communities in which it participates. Intelligence is seen not as the deployment of capabilities for problem-solving, but as the continual, ever-changing and unfinished engagement with the environment in terms of narratives. The construction of the identity of the intelligent individual involves the appropriation or taking up of positions within the narratives in which it participates. Thus, the new approach argues that the functioning mind is shaped by the meaning ascribed to experience, by its situation in the social matrix, and by practices of self and of relationship into which its life is recruited. Classic AI models such as goal-directed problem solving then can be seen as special cases of narrative practices instead of as ontological foundations. There are implications for representation and performativity that have given us new confidence in the form of a 'strong AI' attitude.

1 Introduction

A symposium on designing a functioning mind in the year 2000 reminds me of the endeavours of aircraft inventors about a hundred years ago. Each inventor built his own flying machine and attempted to demonstrate it before witnesses. Every known idea for heavier-than-air flight was represented: machines with flapping wings, machines with a stack of six or more wings, giant rotating helical spirals; airplane/balloon combinations; with power by bicycle, steam and diesel engines. Test flights invariably ended in failure, and whatever achievements were demonstrated, sustained controllable flight was not one of them. We can look at newsreels of early demonstrations with amusement secure in the knowledge that modern air travel is safe and efficient. This is possible because we have only one thing the pioneers of flight lacked: a readily available theory of aerodynamics. The pioneers had sufficient technology and engineering knowledge at the time, but no foundation describing how engineering skills might make the best use of the available technology.

Designing a functioning mind is difficult partly because there are no ready-made theories, and partly because it is not clear that what we can theorise about is relevant to the concern. If there were a reasonably complete theory of the mind, we would be in a position to implement at least parts of it to an impressive degree. And yet this has not yet happened to the satisfaction of most researchers. Other fields have had an easier task because of the existence of ready-made theories developed within the last 150 years. For example, electromagnetism gives the relationship between an electric field and a magnetic field (Heaviside, Maxwell), and can be used to simulate, design and analyse electrical circuits as well as to develop better theories of electromagnetism. Aerodynamics gives a relationship between lift, thrust, drag and gravity. This provides a foundation for numerical simulation of aircraft, helps to design aircraft, helps to understand natural flight, and helps to develop better theories of aerodynamics. AI has a set of ready-made theories such as algorithms for problem solving, but these theories are not about the mind. The lack of a suitable theory of functioning minds is a problem.

Problems are accompanied by paradoxes. The paradox of AI lies in the distinction between mundane and expert tasks. People perform mundane tasks automatically, yet these seem to require complex reasoning. The principles of washing dishes or changing babies' diapers or hammering nails or installing automobile windshields can be taught within minutes to people with widely varying intellectual abilities, creativity and insight. These tasks involve the handling of imprecise quantities of intractable materials such as soap suds, cloth, slippery dishes and glass sheets, all difficult to model mathematically. In these tasks no precise measurement is called for ('squirt about this much soap into the sink'), specifications are incomplete ('then jiggle it round until it fits'; 'then pin these two bits together') and circular ('you know it fits when it snaps into place'), yet dextrous perceptual-motor operations and sensitive judgements are employed. The fact that apparently easy tasks may conceal a large amount of unconscious sophisticated analysis and processing is not the point. Whether these tasks need 'insight' or 'representations' or 'knowledge' or 'problem solving' is also not the point. The point is that these tasks are taught by 'showing how' and learned 'by doing', relying on social and mimetic interaction, with performance parameters limited by embodiment.

By contrast, expert tasks require specialised skill and training, for example medical diagnosis, playing expert chess, financial decision making. Paradoxically, mundane tasks are the harder to automate, and expert tasks are the easier to automate. Why is it hard to program computers to do things people find easy to do? And why is it easy to program computers to do things that people find hard to do? The answer is that implementing expert tasks does not require theories of the person. It requires only a theory of medical diagnosis or chess or financial markets or whatever. We can formulate such theories, and the implemented system can function with reference to and perform relative to such theories, and this is often sufficient to perform the task. On the other hand, implementing the mundane tasks involve knowing how the embodied person relates to things or other persons, and this is one of the theories we lack today.

I have argued elsewhere (Clocksin, 1995, 1998) that AI is limited not by processing power nor in handling complexity, but by taken-for-granted concepts that form an implicit normative load. Historically, AI research has been freighted by a 'command and control' paradigm (Edwards, 1996) in which AI is seen as the application of decision processes, represented using formal reasoning, to abstract well-specified tasks. Of the vast literature demonstrating this point, let two examples (Turner, 1984, Ramsay, 1988) suffice, or look at any large AI textbook. Indeed, the effort to make progress in AI through one style of knowledge representation and automated reasoning, logic programming and Prolog (e.g. Baral and Gelfond, 1994), culminating in Japan's Fifth Generation Programme of the 1980's, is probably the most ambitious and well supported manifestation of this paradigm. Autonomous agent and rational agent research is one of the more recent followers of the 'rationalist' paradigm.

It may be necessary to break away from the rationalist tradition in order to do justice to alternative models of minds, models based on persons, identity and social context, that may ultimately provide a workable foundation for AI achievements. There are signs that this 'reframing' of AI is happening. For example, Picard (1997) argues that AI has long ignored the importance of emotions, and that emotions are part of the essential foundation for intelligence. Although I am in sympathy with Picard's conclusion, I agree with Sloman (1999) that Picard and others misinterpret Damasio's (1994) observations concerning frontal lobe damage. The observed effects of such damage do not actually imply that emotions are essential for intelligence.

The work described in this paper is informed by two strands of thought emerging from social and developmental psychology. First, there has been an increasing concern with personhood: with persons, agency and action, rather than causes, behaviour and objects (Shotter and Gergen, 1989). Second, an emphasis on the self as a social construct, that persons are the result of interactions with significant others, and that the nature of these interactions is in turn shaped by the settings in which these interactions occur (Levine, 1992).

2 Reframing Rationality

While AI's emphasis on logical symbol-processing as the basis of intelligence has been criticised by Weizenbaum (1976) and by Dreyfus (1972), Winograd and Flores (1985) were probably the first to call for a re-examination of the rationalistic tradition in its influence on artificial intelligence research. According to Dreyfus and Dreyfus (1986),

> the failure of AI research will continue until intelligence ceases to be understood as abstract reason and computers cease to be used as reasoning machines.

Furthermore, McDermott (1987) argues that

> ...the skimpy progress observed so far is no accident, and in fact it is going to be very difficult to do much better in the future. The reason is that the unspoken premise..., that a lot of reasoning can be analysed as deductive or approximately deductive, is erroneous. (p 151)

Finally, the following observation from Weizenbaum (1976) has enormous significance:

> ...intelligence manifests itself only relative to specific social and cultural contexts.

Is man, as Aristotle put it, a rational animal? Rationality in logical terms is related to consistency: an argument is rational if it does not produce a contradiction. Beginning with Doyle (1979) and McCarthy (1980), much effort has devoted to methods for preventing contradictions in the presence of changing assumptions. And yet human performance in problem solving is marked by two characteristics that suggest that people sit rather more lightly to logical consistency: people can happily entertain contradictory views (sometimes without being aware of it), and when put to the test, human 'rationality' is frail and fallible. People do not rely on deduction in their everyday thinking. Human experience is marked by incompetence, blunders, and acts of misjudgment. The average AI programmer (but not your average psychologist) might be surprised to learn that for normal people, irrational behaviour is the *norm* rather than the exception. The more accessible background literature on the 'normality' of irrationality includes Sutherland, Manktelow and Over, Dixon, and Piattelli-Palmerini. Human 'reasoning' thrives on the basis of actions and beliefs that cannot be justified nor supported logically.

So is the human mind based on a capacity for rational thought? If so, then why is there a lack of connection between rationality and everyday behaviour? And, if a rational mind is based on a rational brain, then why should we have a rational brain at all, as it seems to find little employment as an engine of rationality? So which is the more likely hypothesis:

> Intelligence derives from a rational mind, operating according to rules of logic (with truth-maintenance and the rest), but is hardly ever used in the way 'intended';

OR

> The mind is not based on principles of rationality, but as people in society we have come to perform and to value a developed but imperfect rational discourse because it has tangible benefits.

3 Construction and the Mind

The social psychology literature tells us that it is useful to make a distinction between ***constructivism*** and ***constructionism***. Constructivism, a term coined probably by Nelson Goodman and also called cognitive constructivism, describes the work of Piaget and Neisser. The main idea is that reality is created through the operation of a variety of mental schemata, analysis and synthesis procedures. This is the main assumption behind cognitive psychology, and it is safe to say this is the prime assumption of AI research. AI is concerned with the design of data structures to implement mental schemata, and the design of algorithms to implement analysis and synthesis procedure.

By contrast, constructionism describes the work of Gergen, Potter and Harré. It has also been called discursive constructivism, which is probably the source of the confusion of terminology. The idea is that constructive processes are to be found in relationships, often discursive, between persons. These relationships embody situated reasoning: that is, a contextualised meaning-producing performance. Figure 1 shows cartoon versions of constructivism and constructionism.

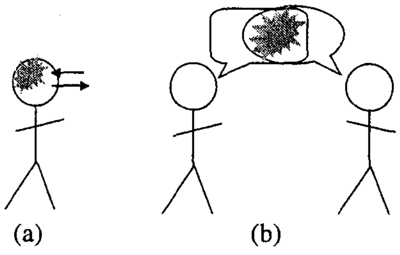


Figure 1. (a) the constructivist concern; (b) the constructionist concern.

The cartoon in Fig 1(a) is a familiar AI model: the problem-solving agent, operating a perception-processing-action loop. Meaning is constructed in the mind as a result of internal capability. By contrast, Fig 1(b) privileges social interaction as the meaning-constructing process. Here the mind is for social development, engagement with persons, and the institution and appropriation of personhood and identity. Figure 2 shows a pair of layers describing the functional dependencies posited by both approaches.

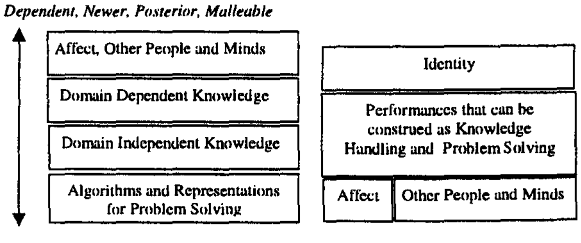

Dependent, Newer, Posterior, Malleable
Affect, Other People and Minds
Domain Dependent Knowledge
Domain Independent Knowledge
Algorithms and Representations for Problem Solving
Identity
Performances that can be construed as Knowledge Handling and Problem Solving
Affect
Other People and Minds
Foundational, Older, Prior, Essential, 'Closer to the hardware'

Figure 2. (a) Conventional layering; (b) Constructionist layering.

An example of way that affect is marginalised by the conventional understanding of the mind is illustrated by Mr Data, the android character on the popular television programme *Star Trek: The Next Generation.* Mr Data can function perfectly well without emotions, but has an optional plug-in 'emotion chip' that makes him more human. With the chip at his disposal, Mr Data is at least more versatile than Mr Spock of the original *Star Trek* series. Half-human, half-Vulcan, Spock is trained to abhor emotion. Given that this is after all science fiction, these characters will particularly appeal to adolescent males who themselves can find difficulty coming to terms with their emotions and the complexities of the social realities in which they are beginning to find themselves. How much more comforting to retreat to a simplified world by identifying with heroes who succeed by applying pure logic and calculation.

By contrast, the constructionist finds such accounts incoherent. Not only are there glaring inconsistencies in the model – the Data and Spock characters actually behave in ways that use the affective skills they are reputed not to require – but also the 'optionality' of affect is associated with individuals one sees only in the human population as profoundly disordered. It is probably intuitions of this type that lead myself, Picard and others to reject the optionality of affect implied in Fig 1(a), and instead to adopt a view like Fig 2(b) which treats the world of affect as something prior to intelligence and problem-solving. Developmentally this reflects the early requirement for social relationship in the form of mother-newborn relationship.

4 The Unfinished Mind

Intelligent behaviour depends on the challenges that arise from being embodied and being expected to participate in social interactions. Every organism is socialised by dealing with -- engaging with, participating with -- the problems that confront it according to its capabilities. This activity has a spiraling effect, as capabilities bring about social experience which in turn modifies -- conditions, improves -- the capabilities for further social experience. The functioning mind's standpoint in this engagement is what brings about -- or institutes or constructs or establishes -- an identity. Because development occurs throughout the organism's life, its identity is never fixed or foreclosed. From the constructionist perspective, identity can never be understood as a prior given, even if the organism may understand it (*i.e.* endorses a metanarrative from a culture that names it) as such. AI research will make progress when this spiral process of social and cultural development, and its relationship to the ontogenesis of the individual's identity, can be understood in clear and unambiguous (*i.e.* computational) terms.

One way to describe the developmental process is the so-called 'hermeneutic circle' of critical theory, depicted in Figure 3.

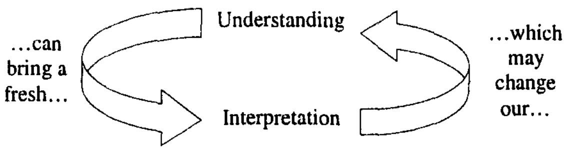


Figure 3. The provisional, contingent, unfinished nature of engagement illustrated by the hermeneutic circle.

Interpretation and understanding influence each other in a circular way. To understand a text, we need to interpret it. But understanding it may lead us to consider a fresh interpretation. When the text is read in the light of the new interpretation, it may change our understanding of it. This may call for a revised interpretation, and so on. This circle is again better described as a spiral, for one never visits the same locus in hermeneutic phase-space (as it were) twice. There is also no reason to restrict the number of trajectories through the space to one. AI research needs to be more attentive to the continual and developmental character of the hermeneutic circle. A system that is not committed to the hermeneutic circle cannot be said to understand. It is this capacity for continual co-evolution (in this example, of interpretation and understanding), or ontogenesis, that is characteristic of a sustaining system. Such a sustaining system is what I would call a functioning mind. The AI thoughtworld needs to reject models such as 'finding the answer to a problem' and 'maintaining consistency of beliefs'. By contrast, ontogenesis of the type advocated here may proceed for a system in which problems and answers are not neatly defined, and consistency in a logical sense is never achievable. Thus there needs to be an emphasis on the *provisionality* of thought, another hallmark of a social constructionist perspective.

5 A Narrative Architecture

The new approach to AI research outlined here is built upon the idea of a narrative, and is adapted from Clocksin (1998). The narrative or story provides a framework that facilitates the interpretation of experience, for it is through the narratives people have about their own lives and the lives of others that they make sense of their experience. Narratives are not restricted to the form of a story with its initial complication followed by its denouement, although stories are good examples of narratives. The idea of narrative is not restricted to written texts. Not only do narratives influence the meaning that people give to experience, they also influence which aspects of experience people select for articulation. Narratives provide not merely a reflection or mirror of life, but provide for the shaping and structure of life. Alasdair MacIntyre (1981) recognises several diverse uses of narrative. He argues that human action is narrative in form, that human life has a fundamentally narrative shape, and that people are story-tellers who position their lives and arguments within narrative histories. Communities and traditions are invested with continuity through narrative histories, and epistemological progress is marked by the construction and reconstruction of more adequate narratives. For Jerome Bruner (1986), narrative is one of the modes of cognitive functioning, a way of knowing that provides a distinctive way of ordering experience and constructing reality. The other mode is logical argument, and Bruner notes that while narrative deals with 'the vicissitudes of human intentions' and is built upon concern for the human condition, logical arguments are either conclusive or inconclusive.

The functioning mind is shaped by a play of images and allusions of the subtle and elusive kind that owes more to the imagination of the storyteller than to the rational operations of the logician. Consider first the question of emplotment. Every story has a plot that involves movement from an initial tension to the resolution of that tension, what Aristotle literally called binding and loosing (*Poetics* 18.1-3) respectively. More generally, a narrative may be described as any time-extended behaviour of a system that exhibits one or more sequential *tension• release* patterns, each of which may recursively embed other *tension•release* patterns within it. Each *tension• release* pattern may be termed an episode. We identify narratives at every scale of time-extended behaviour, from the charge•discharge electro-chemical activity of nerve cells to the impulse•relax rhythmic control of coordinated motor activity, to the expectation•emission of activity that makes a visual or acoustic impact, to the problem•solution of goal-directed behaviour, to the preparation•resolution of musical and poetic cadence, to the structure of songs and dances, building up all the way to people discussing the latest round of tax increases and where they're going to eat lunch. My idea of narrative is informed but not restricted by the ideas of the discursive psychology of Harré and Gillett (1994) who are mainly concerned with higher cognitive functions and who see the mind as

> ...embedded in historical, political, cultural, social, and interpersonal contexts. It is not definable in isolation. And to be a psychological being at all, one must be in possession of some minimal repertoire of the cluster of skills necessary to the management of the discourses into which one may from time to time enter. [p 25-26]

Thus discourse can be related to emplotment, and the other typical dimensions of narrative, namely setting and character, are related to environment (context) and identity (being), respectively.

The main difference between narratives and other structures such as knowledge representations (Way, 1991) is that narratives serve to engage the individual with the contingencies of identity and relationships with others. The individual develops or constructs identity by locating itself within the discourse -- the stories, narratives -- with which it engages. Thus, identity is not simply a matter of building up a store of knowledge or cognitive capabilities, or the self-assembly of numerous micro-personalities into a smaller number of personae (*pace* Jung), but the taking up of a position within a set of discourses as they are negotiated. Therefore, behaviour involves a certain performance and production of a self which is the effect of a discourse, and which has a circular quality. The fact that a developing self may claim to represent itself as a truth prior to the discourse must be seen as a phenomenon emerging from the engagement rather than as an ontological assumption.

This 'taking up of a position' within a narrative may be identified at all levels: from the steady-state behaviour of an electro-chemical process in tissues, to the habituation of a tissue in response to a pattern of conditioning stimuli, to the formation of a reflex. At the highest social level it may involve the articulation of or appropriation of or commitment to a particular policy, a system of beliefs and values, a family group, a football team or a gender role. The ways in which one plays social games is the way in which the identity becomes, to use Butler's (1991) list of words, 'established, instituted, circulated and confirmed' [p. 18]. It is important to remember the ontogenetic nature of the process: the taking up of a position within a narrative is not entirely determined by genetics, nor is it entirely determined by culture. All enabling capabilities participate in their own way with the development of the system.

Let me outline some broad implications of understanding the functioning mind in terms of narrative and the narrative generation process:

a) Narratives can be repeated and imitated. Aristotle defines man as the political and rational animal, but we should also remember his observation that 'man differs from other animals in his greater aptitude for imitation [mimesis]' (*Poetics* 4). Mimesis is seen as the fundamental operator over narratives, implicated in the processes of memory, consciousness and performance described below. Mimesis is not a process of loss-free copying: rather, it a capability that culminates in 're-storying', or re-telling or re-authoring of a story.

b) Narratives, being indeterminate and having hierarchically decomposable emplotment, are a convenient form for concealing and revealing subnarratives. A narrative is extensible both along its length and its hierarchy, for it is always possible to apply questions such as, 'and *then* what happened?', or 'where did he come from?', or 'how did she do that?', or 'why did he say that?', or 'why didn't she do that?' as a means of closing or opening new subnarratives and filling in or creating gaps.

c) Narrative serves as a memory, not only in a collective and institutional sense of telling stories to each other (memorial), but also as the foundation for the individual's memory (Middleton and Edwards, 1990). Memory as narrative provides an alternative to the conventional idea of memory as a storehouse of knowledge and experience, and of remembering as a process of retrieval from the storehouse. Instead, memory is understood not as the storage-and-retrieval manipulation of a network of entities, but rather as a particular practice within the context of a discourse. Memory is an active process in which experience is emplotted (*i.e.* the lived experience is 'recruited' by a prior narrative) and personified (*i.e.* attributed the status of a 'character' in a prior narrative) in ways constrained by the usages and contexts of prior narratives. This 'circular' definition is committed to the idea of memory not as storage and retrieval of facts, but as the continual re-membering of our memories of our memories *(sic)*. Memories cannot be decoupled from the narrative structure of remembering, for outside a narrative they have no meaning (Shotter, 1990).

d) Narratives can be edited, and this is particularly relevant to the question of memory and learning. As Platinga puts it, 'our memories are not inert, but undergo a process of editing, whereby they are regularized, rendered more retainable and reshaped with an eye to subsequent circumstances and events' [p. 45].

e) Learning and therapy are understood as means of editing or rewriting narratives. These both begin with the assumption that people experience problems when the stories of their lives, as they or others have constructed them, do not sufficiently represent their lived experience. Therapy and learning then become processes of storying and re-storying the lives and experiences of people. In this way narrative comes to play a central role in therapy and education, but for the purposes of this paper point to models of 'learning' and 'belief revision' in the functioning mind. There is a significant literature on what might be described in my terms as narrative-theoretic formulations of memorial, mimesis, ritual, education and therapy, for example Elliott (1995), Girard (1977), Buckland (1995) and Epston and White (1992).

f) A narrative is a suitable representation for self-reflection. Whether consciousness is considered as a capacity for, or a result of, self-reflection, consciousness is about making explicit in narrative form (*i.e.* telling a story about) the person's interpretation of the narratives with which it believes itself to be engaged. The nature and extent of the belief is also conditioned by the person's appropriation of narratives from the culture, giving rise to 'cultural conditioning'. Consciousness is the story we tell about where (context) we (character) are and what (plot) we are doing. The existence of what Dennett (1991) calls the 'Cartesian theatre' as a setting for this performance therefore should not be surprising, and certainly does not imply endorsement of Cartesian dualism. Consciousness as an 'account of ourselves' therefore proves necessary because we are animals who partly construct our accounts through interaction with a social world: we need accounts of ourselves because we are accountable 'as selves' to others, and this exchange of accountability again has a spiral instituting character. The nature and purpose of consciousness therefore cannot be understood in relation only to the individual.

g) Narratives can be performed. Performance (whether it has an inward or outward expression) is the means by which narratives are rendered to self and others, and depends on the capacity for mimesis. Performance is more general than reciting a story on a stage. It is the actual means of behaviour. Yet performance, like remembering, does not render a narrative intact. This understanding concurs with Butler (1993, p 187) in opposing the conventional notion that performativity is the efficacious expression of the human will. An act of performance is always ambiguous and open to interpretation. Each performance is a unique act of creation because it opens a new gap between intention and realisation, between presentation and representation.

h) Because mimesis (the mechanism for re-storying, performance, memory) generates non-identical narratives that are open to interpretation, it is a well-spring of perpetual conflict. Conflict leads to rivalry and violence, but also to dialectic and symbolic interaction. Thus the provision of narrative-building tension together with a motivation for social interaction are built in at a fundamental level.

Narratives that emerge at the layer of surface behaviour are particularly important, for all performance is understood within the context of the social group. At a fundamental level, narratives are performed as rituals and myths, both of which also have the instituting effect of maintaining a corporate memory. Myths supply a framework of images, in narrative form, in which the needs, values and fears of the group – in short an articulation of affect within a network of social consciousness – can be expressed (Clocksin, 1995). Rituals are a demonstration of the way that sensations and actions are integrated as a result of their co-development, and are expressed by the articulation of a policy by members of the group. Recalling Weizenbaum's observation that 'intelligence manifests itself only relative to specific social and cultural contexts', I suggest that ritual and myth as the performative means of the social and cultural context are the prior and necessary components for the manifestation of intelligent behaviour. This, like the layerings of Figure 2, implies a view of the mind that is an inversion of the conventional view taken by artificial intelligence and cognitive science researchers. The conventional view assumes that behaviour arises from basic capabilities for cognition and problem solving. The emphasis seems to be on the individual mind that generates solutions to problems. The questions of emotion and interaction with others are considered as additional – possibly even unnecessary or irrelevant – complications or side-issues. By contrast, the 'upside down' view sees problem solving as a practice or policy that has been constructed as a *result* of a basic capacity for social interaction within a cultural context, which itself is based on fundamental reciprocalities of need and desire. Therefore, further progress in artificial intelligence research needs to be attentive to issues arising from the articulation of practices and policies by a group. An example of this is given in Clocksin (1998).

6 Conclusion

I have put forward a view that the functioning mind should not be seen as a kind of abstract puzzle solving applied by an individual to arbitrarily defined problems presented by an alien environment. It is true that certain isolated episodes of intelligent behaviour can be given an interpretation as puzzle-solving, maximising of utility, and so forth. However, we need to accept that there is a close ontogenetic connection between the nervous system and the environment. It is possible to accept that this connection has the effect of individualising precisely which capabilities and information sources are employed by the mind as it develops, without necessarily endorsing a structuralist ontological commitment as to the precise delineations of capability and information source.

The AI system of the future will be based upon what Darwin called 'the social instincts'. The functioning mind cannot be considered in isolation from an embodied existence that is committed to an ontogenetic engagement with the environment. The body matters, and in this way the functions of the mind might be said to resemble physical functions such as digestion which can't take place without a body. But intelligence is of a different order because at its core is the generation and use of images and narratives that serve to circulate within the gap between presentation and representation that is opened up by performance.

I repeat Weizenbaum's dictum: Intelligence manifests itself only relative to a matrix of social and cultural contexts. There is a propinquity of participation in communities that are defined by historically individualised mutualities and reciprocalities of need and desire. The ways we participate with this matrix, and construct identities that take up positions within this matrix, are through policy, contract, and institution. This participation and construction, which makes explicit the political imperative of intelligence, is made possible by a variety of means such as memorial, mimesis, ritual, education and therapy, all of which serve to expose and rewrite narratives. The functioning mind is thus seen as an historical contingency constituted in part by policies articulated by the social group, rather than by the underlying cognitive nature of the individual brain. This perspective on the mind is what Rorty would call 'ironic' and 'political' in the sense that it makes explicit its theoretical commitment to policy, community and institution.

An implication for future research is the idea of 'narrative architectures', built upon a foundation which is attentive to the way that signals and symbols influence (and are influenced by) the way we as communities of individuals make sense of experience, construct our identities, and produce meaning in the world. This makes me optimistic about the prospect for what John Searle called 'strong AI'. The characteristic of an AI system that really can be said to have a functioning mind is that it can perform an account of the identity it has constructed using narratives it has appropriated from the culture and society that it shares with us. Such a system negotiates its positions to become instituted among the group, participating in the social consciousness it constructs with members of the group.

Acknowledgements

I thank John Cornwell, John Searle, Pam Clocksin and Aaron Sloman for discussions.

References

Baral, C. and M. Gelfond, 1994. Logic programming and knowledge representation. *Journal of Logic Programming* **19**, 73-148.

Bruner, J., 1986. *Actual Minds, Possible Worlds*. Harvard University Press.

Buckland, S., 1995. Ritual, bodies and 'cultural memory". *Concilium* 1995/3, 49-56

Butler, J., 1991. Imitation and gender subordination. *Inside/Out*, ed D. Fuss et al. Routledge. 13-31.

Butler, J., 1993. *Bodies that Matter*. Routledge.

W.F. Clocksin, 1995. Knowledge Representation and Myth. Chapter 12 of *Nature's Imagination: The Frontiers of Scientific Vision* (J. Cornwell, ed). Oxford University Press.

W.F. Clocksin, 1998. Artificial Intelligence and Human Identity. Chapter 6 of *Consciousness and Human Identity* (J. Cornwell, ed.) Oxford University Press.

Damasio, A.R., 1994. *Descartes' Error: Emotion, Reason and the Human Brain*. Grosset/Putnam Books.

Dennett, D.C., 1991. *Consciousness Explained*. Penguin.

Dixon, N., 1976. *On the Psychology of Military Incompetence*. Futura.

Dixon, N., 1987. *Our Own Worst Enemy*. Futura.

Doyle, J., 1979. A truth maintenance system. *Artificial Intelligence* **12**, 231-272.

Dreyfus, H.L., 1972. *What Computers Can't Do*. MIT Press.

Dreyfus, H.L. and Dreyfus, S.E., 1986. *Mind over Machine*. Macmillan.

Edwards, P.N., 1996. *The Closed World: Computers and the Politics of Discourse in Cold War America*, MIT Press.

Elliott, C.M., 1995. *Memory and Salvation*, SPCK.

Epston, D. and White, M, 1992. *Experience, Contradiction, Narrative & Imagination*. Dulwich Centre Publications, Adelaide, South Australia.

Girard, R., ET:1977. *Violence and the Sacred*. Baltimore: Johns Hopkins University Press.

Harré, R. and Gillett, G., 1994. *The Discursive Mind.* London: Sage Publications.

Levine, G., 1992. *Constructions of the Self*. Rutgers University Press.

MacIntyre, A., 1981. *After Virtue: A study in moral theory*. Duckworth.

Manktelow and Over, 1990. *Inference and Understanding*. Routledge.

McCarthy, J., 1980. Circumscription – a form of non-monotonic reasoning. *Artificial Intelligence* **13**, 27-39.

McDermott, D., 1987. A critique of pure reason. *Computational Intelligence* **3**, 151-160.

Middleton, D. and D. Edwards, 1990. *Collective Remembering*. Sage.

Piattelli-Palmarini, M., 1994. *Inevitable Illusions: How mistakes of reason rule our minds*. MIT Press.

Picard, R., 1998. *Affective Computing*, MIT Press.

Platinga, T., 1992. *How Memory Shapes Narratives*. Lampeter:Mellen.

Ramsay, A., 1988. *Formal Methods in Artificial Intelligence*. Cambridge University Press.

Rorty, R., 1989. *Contingency, Irony, and Solidarity*. Cambridge University Press.

Shotter, J., 1990. The social construction of remembering and forgetting. In Middleton and Edwards (1990).

Shotter, J. and K.J. Gergen (1989). *Texts of Identity*. Sage.

Sloman, A., 1999. Review of Affective Computing, *AI Magazine* **20**(1), 127-133.

Sutherland, S., 1992. *Irrationality: The enemy within*. Penguin.

Turner, R., 1984. *Logics for Artificial Intelligence*. Ellis-Horwood.

Turner, R., 1990. *Truth and Modality for Knowledge Representation*. Ellis-Horwood.

Way, E.C., 1991. *Knowledge Representation and Metaphor*. Kluwer.

Weizenbaum, J., 1976. *Computer Power and Human Reason*. Penguin.

Winograd, T. and Flores, F., 1985. *Understanding Computers and Cognition: A new foundation for design.* Addison-Wesley.

Minds have personalities - Emotion is the core

Darryl N. Davis

Neural, Emergent and Agent Technology Research Group,

Department of Computer Science, University of Hull,

Kingston-upon-Hull, HU6 7RX, U.K.

D.N.Davis@dcs.hull.ac.uk

Abstract

There are many models of mind, and many different exemplars of agent architectures. Some models of mind map onto computational designs and some agent architectures are capable of supporting different models of mind. Many agent architectures are competency-based designs related to tasks in specific domains. The more general frameworks map across tasks and domains. These types of agent architectures are capable of many cognitive competencies associated with a functioning mind. However, there is a problem with many of these approaches when they are applied to the design of a mind analogous in type to the human mind – there is no core other than an information processing architecture. As any specific architecture is applied to different domains, the information processing content (knowledge and behaviours) of the architecture changes wholesale. From the perspective of developing intelligent computational systems this is more than acceptable. From the perspective of developing or simulating functioning (human-like) minds this is problematic – these models are in effect autistic. This paper presents an emotion-based core for mind. This work draws on evidence from neuroscience, philosophy and psychology. As an agent monitors its internal interactions and relates these to tasks in its external environment, the impetus for change within itself (i.e. a need to learn) is manifested as an unwanted combination of emotions (a disequilibrium). The internal landscape of emotion, control states and dispositions provides a basis for a computational model of personality (and consciousness).

1. Introduction

For much of its history, cognitive science has positioned emotion as the poor relation to cognition. For many emotion is the Achilles' heel of reason. This paper takes a stance on (human-like) minds that places emotion as the core. From a computational perspective, the impetus for this research is the inadequacy of earlier work on the modelling of motivation (Davis 1996) to adequately contain aspects of cognitive functioning. This paper takes a trajectory through work from neuroscience on what parts of the central nervous system play a role in emotions, research from psychology and analyses from philosophy. This paper will not give a definitive definition of emotion but look to argument and finding agreement from a number of sources in ascribing the role of emotion in functioning minds. A sketch of a computational theory of mind (primarily from the agent perspective) will be then be considered in the light of this evidence. This leads onto the presentation of preliminary experimental work that models emotions as the core of a computational architecture of a mind.

2. Emotions and the mind

The nature of emotions and the relation to thought have been analysed since the dawn of western civilisation. Plato degrades them as distorting rationality. Aristotle denotes long tracts to their categorisation and impact on social life. For Darwin emotions in adult humans are a by-product of evolutionary history and personal development.

Here the definition of emotions as "*...examples of non-problem-solving non-behaviour*" (Gunderson 1985:72) is completely rejected. Merleau-Ponty supposes humans are moved to action by disequilibria between the self and the world. Emotion plays a large role in initiating and providing descriptors for such disequilibria. Emotion is a primary source of motivation. Criminal law recognises the importance of emotions in differentiating between voluntary manslaughter (occurring in the heat of passion) and murder (involving malice aforethought and deliberate suspension of control). French law takes this further with its concept of crimes of passion. However to consider emotions solely as an emergent quality of mental life that undermines reason and rationality is "*a vehicle of irresponsibility, a way of absolving oneself from those fits of sensitivity and foolishness that constitute the most important aspects of our lives*" (Solomon 1993:131-132). Emotions are "*a subjective strategy for the maximisation of personal dignity and self-esteem*" (Solomon 1993:222). Schenck (2000) in his study of the role of music suggests that there are resource and motivation problems associated with this tension between emotions and cognition and that "*we are rational only when we have the time, or the inclination to be so*". Much of psychopathology and psychiatry is concerned with understanding how minds dysfunction. Depression, mania and phobias are often associated with affective disorders. Much of the treatment of depression revolves around identifying and correcting the sources for the emotions of fear and anxiety. Damasio's text (1994) details how physiological damage to the prefrontal cortex, the limbic system (in particular the amygdala) and the afferent pathways that connect the two areas result in emotional dysfunction, personality change and a loss of reason (dissociation). Again emotions play an important role in the executive aspects of cognition, i.e. judgement, planning and social conduct. Goleman (1995) terms this emotional intelligence - it appears to be very similar to what others (see Spaulding 1994) term social intelligence. Emotion has many functions including the valencing of thoughts related to emerging problems, tasks and challenges in terms of emotional intensity and emotion type, as in for example directing attention to aspects of internal and external environments. Such a function is a precursor to problem solving. Many researchers have written on the importance of emotion for motivation (Simon 1979; Spaulding 1994), memory (Rolls 1999), reason (Damasio 1994) and learning. Solomon suggests that "*there is no ultimate distinction between reason and passion*", and that together the two provide more than an understanding of experience, they constitute it. In short emotion has a central role in a functioning mind.

The conjecture cognitive scientists need to face is whether the computational modelling of human-like minds is possible without a silicon/digital analogue to human-like emotions. Research into producing computational cognition may lead to the development of intelligent problem-solvers of many types (e.g. ACT, AIS, SOAR), but the simulation of the human mind requires other categories of intellectual processes. Much of cognitive science and artificial intelligence adopts a modular approach to cognition. If vision, memory, attention, language can be solved, an artificial brain can be built. Such an artefact will perceive, reason and act in its world, relating current to past events, focusing on cognitive salient events in that world. It will interact with and represent parts of its external environment but it will have no internal environment and no sense of self. Without emotions it will be diagnosed as autistic! This approach to cognitive science is one that Harré (1994) argues against - the individual as passive observer of the computational processing that is that person's cognition. Cognition is part of the mental repertoire – perhaps a large part but it is not the entirety of the mind. The efficacy of its use depends on the mind it serves. In looking for general principles to the functioning of mind, cognitive science has perhaps neglected those aspects of mental life that give rise to individual differences. This is perhaps understandable as science looks to general principles. However a redress is called for, and to understand how a mind functions, general principles that also explain individual differences need to be found.

An alternative stance is to place emotion at the core of mind. This core gives rise to episodic states (e.g. feelings), trajectory states (e.g. moods and dispositions) and (semi-permanent) endogenous states (e.g. personality). Personality traits lasting years (or a lifetime) are usually tightly bound to qualities of emotions. To rephrase a previous revolution in artificial intelligence: *human-like intelligence requires embodiment of the supporting computational infrastructure not only in terms of an external environment but also in terms of an internal (emotional) environment.*

3. Psychology and emotion

Over the last hundred years of psychology (from James onwards) the study of emotion has waxed and waned with theories of emotion typically rooted in discussions of physiological and non-rational impulses and drives. An exception is the "cognitive" school of emotion dating from Paulhan (1887) through to Schacter and Singer's (1962) influential experiments with adrenaline and the effect of social context on emotive appraisal. A standard introduction to psychology from the 1970s (Lindsay and Norman 1972) summarises much of the experimental work on emotions in suggesting that emotional states are manipulable through cognitive processes (in particular expectations), physiological states and environmental factors. They conclude that cognition (particularly memory, motivation, attention and learning) and emotions are intimately related. In Newell's seminal work on cognition (Newell 1990), emotion is not indexed and is only discussed in any length in relation to social aspects of a cognitive agent in the final chapter. Although Newell acknowledges this, it reflects a trend in cognitive science to place emotion as subordinate to rationality and cognition. Despite pointers to the importance of understanding emotion for cognitive science (e.g. Norman 1985), cognitive science all too readily follows as a modern day Stoic successor to Plato in minimising the role of emotion. A leading volume on the dynamics approach to cognition (Port and Van Gelder, 1995) is no exception – particularly odd if emotion is viewed as the *flow and change* of cognitive predisposition over time and across occasion (Lazurus 1991).

Ortony et al (1988) consider cognition to be the source of emotion, but that unlike many other cognitive processes, emotions are accompanied by visceral and expressive manifestations. They consider valence (i.e. positive-neutral-negative) and appraisal (cognitive reflection of these valences) as the primary basis for describing an emotion. They differentiate emotions from non-emotions on the basis of whether a valenced reaction is necessary for that state. However, non-emotion states (e.g. abandonment) can give rise to causal chains of emotive reactions leading to highly valenced (emotive) states. They suggest that there are basic classes of emotion related to valenced states focussed on events (pleased vs. displeased), agents (approving vs. disapproving) and objects (liking vs. disliking). Specific emotions are instances and blends of these types and subclasses. Emotions of the same type have eliciting conditions that are structurally related. They reject the idea of emotions such as anger and fear being fundamental or basic emotions. The cognitive processing that appraises emotions is goal-based and resembles the type of processing and structures discussed in motivation for autonomous agents (e.g. Beaudoin and Sloman 1993, Davis 1996).

Oatley and Jenkins (1996) define emotion as "*a state usually caused by an event of importance to the subject. It typically includes (a) a conscious mental state with a recognizable quality of feeling and directed towards some object, (b) a bodily perturbation of some kind, (c) recognisable expressions of the face, tone of voice, and gesture (d) a readiness for certain kinds of action*". Others (e.g. Frijda 1986) give similar definitions. A number of other psychologists (e.g. Power and Dalgleish 1997) appear to be in agreement in defining what are basic emotions:

- Fear defined as the physical or social threat to self, or a valued role or goal.
- Anger defined as the blocking or frustrations of a role or goal through the perceived actions of another agent.
- Disgust defined as the elimination or distancing from person, object, or idea repulsive to self and to valued roles and goals.
- Sadness defined as the loss or failure (actual or possible) of a valued role or goal.
- Happiness defined as the successful move towards or completion of a valued role or goal.

They suggest that these five suffice as the basic emotions as they are physiologically, expressively and semantically distinct. There are cases for other emotions to be considered as further basic emotions. From a perspective of classifying emotions using distinctive universal signals, i.e. expressions (Ekman & Davidson 1994), surprise is included in this fundamental set. However, from the perspective of classifying emotions based on distinctive physiological signs (see Power and Dalgleish 1997), the basic set is reduced to fear, anger, disgust and sadness.

Rolls (1999) presents a different perspective on the psychology of the emotions. Brains are designed around reward and punishment (reinforcer) evaluation systems. While this can be seen as analogous to the valenced arousal states in the Ortony et al. theory, the reinforcers are precursors to any specific emotion. Rather than reinforcing particular behavioural patterns of responses (behaviourism), the reinforcement mechanisms work in terms of cognitive activity such as goals and motivation. Emotions are states elicited by reinforcers. These states are positive when concerns (goals) are advanced and negative when impeded. Again, there is an overlap with the perspectives of Power and Dalgleish, and Oatley and Jenkins. These states are more encompassing than those states associated with the mere feelings of emotion. This aspect is considered further in Wollheim's analysis of the emotions. Emotions have many functions (Rolls lists ten) including the priming of reflexive behaviors associated with the autonomic and endocrine system, the establishment of motivational states, the facilitation of memory processing (storage and control) and maintenance of the "*persistent and continuing motivation and direction of behavior*". In effect Rolls suggests that the neuropsychological evidence supports the conjecture that emotions provide the glue that binds the multitude functions of mind.

4. Philosophy and emotion

Wollheim (1999) distinguishes two aspects of mental life in his analysis of emotion: the phenomena of mental states and mental dispositions. Mental states are temporally local to their initiating event and transient, being relatively short-lived - sometimes instantaneous. Mental states can reoccur frequently to give the impression of a continuous state. Mental dispositions can more long-lived (sometimes over a lifetime) – they are temporally global - they have histories. Mental states and dispositions are causally related. Mental states can instantiate and terminate mental dispositions.

Mental states can reinforce and attenuate mental dispositions. Mental dispositions can also facilitate mental states. Both mental states and dispositions have a psychological reality. Impulses, perceptions, imaginings and drives are mental states. Beliefs, knowledge, memories, abilities, phobias and obsessions are examples of mental dispositions. Three very general properties characterise these two types of mental phenomena: intentionality, subjectivity and three exclusive grades of consciousness (conscious, preconscious and unconscious). Both mental states and dispositions have an intentional quality – i.e. they are related or directed to either internal or external events. Wollheim suggests that subjectivity be only associated with mental states – mental dispositions can only be indirectly experienced through the mental states in which they are manifest. It is in highlighting the very differences between mental states and dispositions that Wollheim makes use of the emotions. Emotional states differ from emotional dispositions. Emotions are preconscious mental dispositions and cannot be directly experienced. What can be experienced are feelings or perceptions of emotion (mental states) associated with mental dispositions. While the two can be causally interrelated this need not be the case. Mental dispositions are preconscious (and in some cases unconscious) traits. We can become aware of (aspects of) them though training (e.g. yoga) or therapy and in doing so make parts of the preconscious mind conscious. In everyday functioning however the conscious mind is aware of mental states and relates these to personal histories and intended futures – the current, past and intended states of being.

From an computational perspective on the philosophy of mind, Sloman has for many years considered that intelligent machines will necessarily experience emotion (-like) states (Sloman and Croucher 1987). Following on from the work of Simon (1979), his developing theory of mind and the nature of problem solving considers how in attempting to achieve multiple goals (or motivators) perturbant (emotion-like) states ensue (Wright et al 1996). These perturbant states will arise in any information processing infrastructure where there are insufficient resources to satisfy current and prospective goals. Sloman (1987) tends to describe emotion in terms of disturbances of mental processes (the Achilles heel again!). Like Wollheim, Sloman differentiates between episodic and persistent mental phenomena, both of which can carry emotional constituents. More recently his architectures for functioning minds include primary, secondary and tertiary emotions (Sloman 1999). Primary emotions are analogous to arousal processes in the theories introduced above (i.e. they have a reactive basis). Secondary emotions are those initiated by appraisal mechanisms (i.e. they have a deliberative basis). Tertiary emotions are cognitive perturbances - negatively valenced emergent states - arising from (typically goal or motivator) conflicts in an information processing architecture. Tertiary emotions arise from the interaction of emotions and other cognitive processes (e.g. motivation) at the deliberative layer. In many situations these perturbant states arise through resource inadequacy or mismanagement while pursuing multiple and not necessarily incompatible goals. While the work that follows certainly builds upon some of these ideas, this framework seems flawed. Perhaps the differentiation that Sloman makes between these emotions can be more easily explained in terms of the different categories of processing that the mind performs over its different layers. A secondary emotion is an analogous state (or disposition) to a primary emotion but seemingly perceived in a different manner due to the characteristics of the processing at the different layers. In visual perception terms, the red object that swept past our visual senses, causing a startled (reactive) response, that disturbs ongoing thought and behaviour patterns, is the same red object that is subsequently perceived as a rose petal blown by the wind from a nearby shrub in the garden.

5. Theoretical Framework

The theoretical framework presented here builds on those aspects of agreement in the work presented above. It revisits an earlier (computational) architecture of mind and emphasises the interplay of cognition and emotion through appraisal, motivation and niche space. Psychological definitions of emotion have been presented that refer to cognitive (appraisal) and physiological factors (arousal), and the valencing of emotive states and reinforcers as precursors to emotional arousal. The processes leading to the experience of emotions (in humans) are neither bottom-up nor top-down – they are both and more. Emotions are experienced as a result of the interactions within and with a synergistic information processing architecture that includes (at least) the endocrine system, the limbic system, and the cortices. Emotions, in socio-biological agents, are in part mental (appraisal) states and supporting (valencing) and causal (reinforcer) processes. Any computational model of emotion must attempt to meet similar specifications, and address the differentiation in mental phenomena that Wollheim makes. In moving towards a model of emotion that will be computationally tractable, the extent of the model will be initially (at least) minimised. A minimal model of emotion enables the model to be used as the core to an agent-based model of the mind.

Earlier research on agents focussed on an architecture that supports motivation (Davis 1996). The architecture (sketched in figure 1) emphasises four distinct processing layers: a reflexive layer that is analogous to the autonomic systems in biological agents, a reactive (preconscious) layer, a deliberative layer and a reflective layer. This broad picture of the mind has high level and low level processes co-existing and interacting in a holistic manner. Hence motivator processing, planning, decision-making and other cognitive processes are not merely abstract but exist in relation to other automatic, autonomous and adaptive processes. The entirety of the agent's processing exists in relation to the agent's environmental stance; i.e. what objects, agents and events are occurring in the environment and how they affect the goals and motivations of the agent. The two lower layers relate to *pre-attentive* processes and are capable of supporting innate and learnt environmental competencies and (internal and external) behaviours. Perception of and action upon the external environment is mediated primarily through these two layers. The third (deliberative) layer relates to the types of things discussed in most cognitive science, for example (Newell 1990). This does not preclude a non-symbolic implementation of this layer. The fourth layer, the reflective qualities, serves to monitor the overall behaviour of the agent. In particular, the role of the reflective layer is to identify and act on out-of-control behaviours, whether internal, external, deliberative or reactive. This (reactive, non-deliberative) meta-management level processing is considered to be the most abstract level of processing. If it were not, there is a requirement for the reflective processes to be monitored in turn - this in effect would lead to an infinite regress.

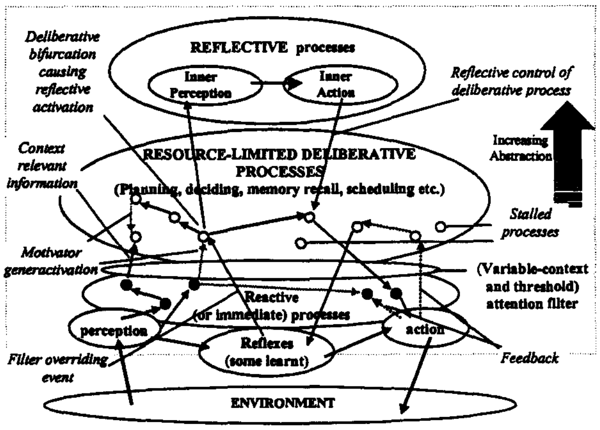

Deliberative bifurcation causing reflective activation
REFLECTIVE processes
Inner Perception
Inner Action
Reflective control of deliberative process
Context relevant information
RESOURCE-LIMITED DELIBERATIVE PROCESSES
(Planning, deciding, memory recall, scheduling etc.)
Increasing Abstraction
Stalled processes
Motivator generactivation
(Variable-context and threshold) attention filter
Reactive (or immediate) processes
perception
action
Reflexes (some learnt)
Filter overriding event
Feedback
ENVIRONMENT

Figure 1. Sketch of an architecture of a mind.

Control suggestions from reflective layer do not always override processes originating and ongoing in the other layers. The behaviour of an intelligent cognitive agent is not controlled by any of these layers in isolation. Behaviours at the reactive level may preclude processes at or actions motivated by the deliberative or reflective layers. Processes over any specific combination of layers may arise as a result of an agent attempting to manage control states originating in any of the layers. Where decision processes related to possibly antagonistic behaviours are not cleanly integrated, there is the very real possibility that the agent will experience cognitive perturbance, particularly where the underlying motives are acute (Wright et al 1996). This cognitive perturbance can be described within an emotional context using tertiary emotions (Sloman 1999).

This analysis presents an incomplete picture. In the earlier work the primary analysis of the mind and the resulting computational designs and systems focussed on motivation and goal processing. This analysis was phrased in terms of niche spaces, design spaces and control states. The niche-design space analysis is still valuable tool in designing a functioning mind. However Wollheim's analysis of the mind and emotions, if accepted, will ultimately require a review of the taxonomy used to relate different control states. A deeper analysis of these control states is required, in terms of temporal extent, subjectivity and grades of consciousness. The structures used in modelling motivation incorporated an emotional indicator that corresponds to a deliberative analysis of the motivator and its context. This semantic labelling is insufficient to model emotions. In biological agents emotions are experienced in a conscious, preconscious and physiological sense, and to some lesser or greater extent in terms of post-hoc rationalisation. Over a lifetime, given no cerebral dysfunction, this emotional landscape is navigated in the attempt to achieve life-goals. This can be viewed as moving between neighbouring niche spaces – for example in moving from music student to professional musician. More dramatic changes in desired niche-space are obviously possible. Different trajectories (goal-achieving behaviours) are possible for any such move. Some trajectories while impossible are supported or attended to for any number of reasons. Emotional intensity associated with the preferred niche space (as in the case of grief and the loss of a loved one) is one example. The preferred trajectory between these niche spaces depends on personality and preferred aspects of the emotional landscapes. The emotional landscape is our internal niche space that allows us as biological agents to understand external events, objects and agents in terms of internal (personal) experience. Our biological design (and psychological capabilities and preferences) define the constraints that determine whether any trajectory between niche spaces is possible (or desired).

The emotional landscape that needs to be modelled in building a functioning mind has to address the four layers of the architecture. Figures 2 and 3 present an integrated model of emotion at the core of a simplified version of the architecture given in figure 1. This model is built upon a trajectory through the research presented in the first half of this paper. An agent typically maintains an ongoing (globally temporal) disposition.

The nature of this disposition is (perhaps only temporarily) modified through current goals and motivations. Over time events occur that modify, stall, negate or achieve goals. Such events can occur over all layers of the architecture. These events give rise to reinforcers. The emotion(s) they reinforce depends on their interactions with conscious and preconscious states and dispositions. A valencing component is needed for any emotion. Both the reinforcer and the (preconscious) valences can be modelled using the interval [-1,1] - this interval need not be linear. A discrete version of this interval maps onto the three tokens: negative, neutral and positive. Thirst, hunger, reproduction etc. are physiological and genetic drives, not emotions. These can be associated with reinforcers and be valenced. They can also be associated with motivators – not all motivators need a source in the emotions. The management and success (or otherwise) of these drive-generated motivations can give rise to emotions. There is case for basic emotions. There is considerable agreement that the set of basic emotions includes anger, fear, disgust and sadness. The definitions given above suffice with one exception. Sadness and happiness are antipathetic, being reflections of each other, or extremes on one dimension. Here the term sobriety is used, with sadness and happiness either side of a neutral state. Sobriety is then defined as no change to a valued role or goal. Happiness and sadness are defined as above. A salient feature of the Oatley, Jennings, Power and Dalgleish definitions of emotion is that they are described in terms of goals, roles and expressive behaviours. This enables emotions to be defined over different levels of the architecture using different aspects of motivational behaviours. The type and subclass analysis of Ortony et al. can be used to build upon this basic set of emotions. The resulting four dimensional model is computationally tractable, and maps onto our ideas for the types of cognitive processing (with particular regard to motivation) that occurs in a mind.

Emotional events are temporally short, although emotional states resulting from successive waves of emotional events can be more enduring. Emotions can be casually inter-related and cause other events. Drives and motivations are highly inter-linked with emotions. These can be embodied in some representation (not necessarily semantic) and in effect relate short-term emotive states to temporally global processes. It is suggested that personality traits are focused at the reflective layer and permeate the rest of the architecture, providing the control patterns that stabilise a personality. Personality traits can be seen as dispositions that affect the reflective processes and influence the different categories of cognitive and animated behaviour. Personality becomes an emergent property of the entire architecture and its disposition to favour specific aspects of the possible emotional landscape, and concentrate on tasks that maximise that aspect of the landscape. Personality traits affect and influence the different categories of cognitive and animated behaviour. Moods arise from the interaction of current temporally global niche roles (the favouring of certain aspects of emotion space) and temporally local drives that reflect the current focus of the deliberative processing. Temporally-global drives are those associated with the agent's overall purpose related to its current, possible and desired niche spaces. Temporally-local drives are related to ephemeral states or events within the agent's environment or itself. These can give rise to more enduring motivational states, which may be acted on.

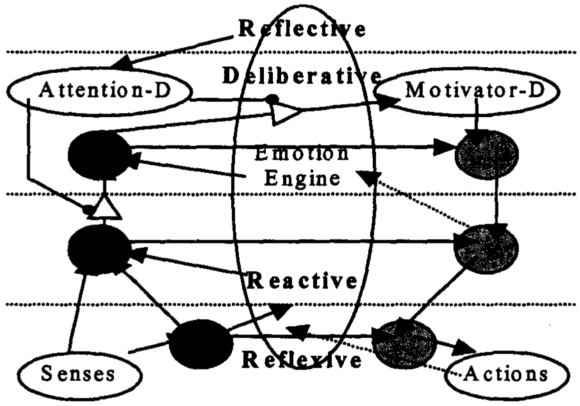


Figure 2. Sketch of the simplified four-layer architecture with emotion as the core. Dark grey circles represent information assimilation and synthesis processes. Light grey circles represent information generation processes that typically mapping into internal and external behaviours. White triangles are filters. Dashed lines represent feedback.

If emergent behaviours (related to emotions) are to be recognised and managed then there must be a design synergy across the different layers of the architecture. Processes at the deliberative level can reason about emergent states elsewhere in the architecture using explicit representations. The reflective processes can classify the processing patterns of the agent in terms of combinations of the four emotions and favoured emotional dispositions. The emotion-changing (reactive) behaviours can be used to pursue a change in emotional disposition. However emotion is not purely top-down processing – as highlighted by Solomon in his differentiation between passion and emotion. Aspects of emotions can be preconscious and, for example, be managed by the autonomic nervous system and its biological substrate (including the endocrine systems). Emotions can move into the conscious mind or be invoked at that level (through cognitive appraisal of agent, object or event related scenarios). Emotions can be instantiated by events both internal and external at a number of levels of abstraction, whether primary (genetic and/or ecological drives), behavioural or by events that require substantive cognitive processing. In the model in figure 3, intense emotions effectively override the emotion filter causing the forced deliberative consideration of the emotional state. Similar filters are used in the earlier work on motivator generactivation (Davis 1996). The deliberative appraisal of the emotion then acts laterally at the deliberative layer, affecting memory management, attention filters and motivator management. The reflective layer of the mind, which is described entirely in terms of the emotion engine, responds asynchronously to the deliberative phenomena.

6. Experimental computational work

The architecture for a computational mind is based on ideas developed within the Cognition and Affect group at Birmingham (Beaudoin and Sloman 1993; Davis 1996). Rather than reiterate the computational work on the non-emotion aspect of that architecture, here preliminary computational and design experiments with the emotion engine are presented.

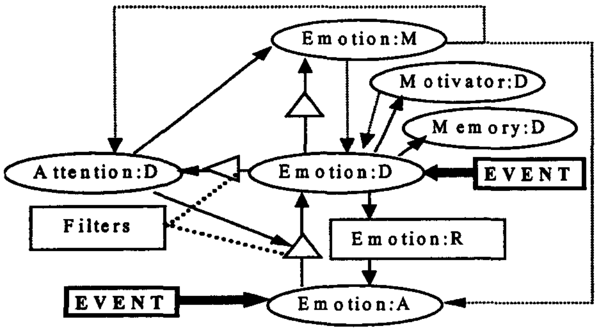


Figure 3. The Emotion Engine for figure 2.

Figure 3 presents a four layer processing model of the emotions. The autonomic processes (Emotion:A) present a base for the model both for dispositional processing and inflection of ongoing dispositions through preconscious events. Such inflections are instantiated by events both external and internal to the agent. The reactive behaviours (Emotion:R) control the functioning of all the preconscious processes. The currently extant Emotion:R behaviours are set by deliberative processes (Emotion:D). The Emotion:M module encompasses the entirety of the meta-management (reflective) processes in this model of the mind. The reflective processes monitor the deliberative appraisal of the Emotion:A processes and the state of the attention filters (managed by Attention:D). The output from Emotion:M provides guidance to the attention management, Emotion:D and the Emotion:A processes. The agent learns to manage its emotions through the development of these five modules. Other aspects of the emotion engine are the placement of deliberative motivator processes, directly affected by Emotion:D. Memory management (Memory:D) is similarly affected.

For a number of reasons the Emotion:A module is modelled using multiple communities of cellular automata. This builds on earlier work (Davis et al 1999) in landscaping decision spaces for the game of Go, and the usefulness of using cellular automata for the modelling of complex social dynamics (Hegselmann and Flache 1998). The behaviours associated with the Emotion:R module govern the internal behaviour of single cells, the communication between adjoining cells in communities and inter-community communication. Different community types have been used. The first experiments (Davis 2000) made use of an insect hive metaphor, with each hive representing an (preconscious) emotional disposition. At the centre of the hive is a (four-dimensional) queen cell that represents the four basic emotions (anger, disgust, fear and sobriety). Each dimension is discretely valenced as positive-neutral-negative. Surrounding the queen cell are four (3-state) drone cells; each mirroring one of the emotions. The remaining (2-state) cells act as filters (guards) or information carriers (worker cells). Further CA communities are being experimented with. The other community type (mobiles) consists of guard and drone cells. This community type represents a reinforcer - a valenced pre-emotive event. Communication between different hives (and input from events outside of the emotion engine at the preconscious level) is by means of the mobile communities. The behaviour of each cell and inter-cell communication is governed by 10 sets of behaviours (50 behaviours in total) plus another behaviour set for inter-community communication. The currently set behaviour from these eleven sets for any hive or hive-mobile combination is selected (as a reactive disposition) by a deliberative (Emotion:D) process. These processes are also responsible for asynchronously monitoring these communities in response to intense hive states and to guidance from the meta-management (Emotion:R) module. Experiments have shown that from any given state, the CA communities rapidly achieved a steady state. By changing the currently extant behaviour set or by communicating with another hive (through the use of a mobile) transitions to the same or other steady states occurs. The CA communities are therefore capable of representing transient and persistent dispositions. The deliberative processes change their emotional disposition (the temporally-global aspect of emotions) and hence the currently extant behaviour set for their hive in response to the reflective processes. The deliberative processes also disturb the motivator and the attention management processes as part of the emotive state appraisal mechanism. Appraisal occurs in response to highly valenced emotive states at the CA communities, feedback from the motivation module or from events occurring elsewhere in the global architecture at the deliberative level. Memory management (Memory:D) also responds to the Emotion:D processes in order to provide emotional context to the storage of memories about external events, objects and agents. The attention filter processes also make use of the state of Emotion:D-Emotion:A complexes to provide a semantic context for motivator filters. The quantitative emotion filters in figure 3 are set directly by the Attention:D mechanism. The intensity levels of these filters are set in response to the Emotion:D mechanisms and the reflective component of the emotion engine.

Learning in the emotion engine occurs in two ways. The reflective mechanism is being implemented using a recurrent neural network that reflects the CA hive communities. Training of the network is given in terms of preferred states within the overall emotional landscape of the cellular automata communities. Further work will look at other types of neural architectures for this and other parts of the emotion engine. The other learning mechanism is the development of preferred reactive behaviour (Emotion:R) combinations in the Emotion:D processes for a particular transition between the steady states of the Emotion:A communities. This is seen as an adaptation of the emotion engine in toto. Currently an experimental harness is being developed, using the Sim_Agent toolkit (Davis et al 1995), in which the emotion engine is trained to prefer specific combinations of emotions, for example the four emotions in similar valences (i.e. all negative, positive or neutral). Artificial scenarios are then provided in which the hive(s) are set in specific or random configurations. As different "personalities" prefer different aspects of the emotional landscape, the engine modifies itself so that preferred emotional states arise as valenced events occur, and preferred dispositions are maintained over longer time spans. Once satisfied that this framework is performing as expected, the earlier motivational architecture will be redesigned to incorporate the emotion engine. This will allow experimentation with emotionally-valenced motivators and allow the investigation of the referenced research using a deeper model of computational mind.

7. Future work

The primary reason for the *preliminary* research described above was to gain a better understanding of the relations between emotion, cognition and mind. Although earlier research on the computational modelling of motivation looked promising, there was a psychological implausibility with the motives behind motivators. Events in an agent's external environment can be represented in terms of motivational descriptors that connect the internal and external environments. The events in an agent's internal environments are described in terms of a synergy over different categories of (internal) computational processes that relates emotions, moods, personality and control. This paper places emotion at the core of the mind. This is analogous to the radioactive cores at the centre of a thermo-nuclear power plant. The plant needs those cores to function but they are not the full story to the functioning of the plant. If synthetic agents are going to experience emotions because of the nature of multiple-goal processing, then the computational infrastructure of those agents needs a representational framework in which these emergent qualities could be harnessed. The emotion engine is one small step in that direction.

While the described work may (superficially) satisfy Picard's (1997) five components for an agent experiencing emotions, the preliminary work is incomplete in a number of ways. The interplay of the reflective and reflexive components requires considerable more work. Preliminary experiments using MLP networks for the reflective processes proved unacceptable at the design stage. Current investigations look to mechanisms that move between discrete (three) space and the non-linear interval, with the queen-cells of currently active hives mirrored in the reflective network. This mechanism also needs to select the appropriate reactive (Emotion:R) behaviours for the preferred combination of emotional dispositions. A more sophisticated architecture would accept non-preferred emotional dispositions in order to achieve important (but temporally local) goals. Preferred dispositions are made non-extant while these goals are achieved. This is an issue that will need to wait until the emotion engine is placed within the architecture shown in figures 1 and 2. Then comparisons with other computational models of emotion, for example (Velásquez 1998) will be possible. Further analysis and investigation will determine whether it is possible to categorise emotion combinations in a manner analogous to the Ortony et al analysis. The discrete version of the basic set of emotions means there are at least 80 possible combinations of emotions; more if event, object and agent directed subtypes are considered. This paper has purposely ignored the social context for emotions, on which there is considerable study from Aristotle to today (see Elster 1999). This is a further inadequacy of the computational theory sketched here.

It has not been possible to review all pertinent evidence within the remit of this paper. The research into the nature of consciousness, and how it might be accomplished within a computational framework, has been glossed over. We accept Wollheim's differentiation between conscious, preconscious and unconscious mental states, and reiterate that any theory that underplays the role of emotions (and personality) in this and other mental phenomena is seriously flawed, as suggested by over 100 years of neuroscientific, psychological and psychiatric evidence. Two of Wollheim's three levels of consciousness map onto the computational framework of reflexive, reactive, deliberative and reflective processes – the theory and model have yet to incorporate the unconscious. It remains unclear whether this will enable a computational agent to experience emotion in the same sense that biological agents experience emotion.

8. References

Beaudoin, L.P. and A. Sloman, A study of motive processing and attention, In: *Prospects for Artificial Intelligence,* Sloman, Hogg, Humphreys, Partridge and Ramsay (Editors), IOS Press, 1993.

Damasio, A.R. *Descartes' Error : Emotion, Reason and the Human Brain,* MacMillan Books, 1994.

Davis, D.N., Sloman, A. and Poli, R. *Simulating agents and their environments.* AISB Quarterly, 1995.

Davis, D.N., Reactive and motivational agents. In: *Intelligent Agents III,* J.P. Muller, M.J. Wooldridge & N.R. Jennings (Editors), Springer-Verlag, 1996.

Davis, D.N., T. Chalabi and B. Berbank-Green, Towards an architecture for artificial life agents: II, In: M. Mohammadian (Editor), *New Frontiers in Computational Intelligence and Its Applications,* ISO Press, 1999.

Davis, D.N., Modelling emotion in computational agents, *Paper submitted to ECAI2000,* 2000.

Ekman, P. and R.J. Davidson (Editors), *The Nature of Emotion,* Oxford University Press, 1994.

Elster, J., *Alchemies of the Mind: Rationality and the Emotions,* Cambridge University Press, 1999.

Frijda, N., *The Emotions,* Cambridge University Press 1986.

Goleman, D.P, *Emotional Intelligence,* Bloomsbury Publishing, 1995.

Gunderson, K., *Mentality and Machines* (2e), Croom Helm, 1985.

Harré, R., Emotion and memory: the second cognitive revolution. In: *Philosophy, Psychology and Psychiatry,* A.P. Griffiths (Editor), Cambridge University Press, 1994.

Hegselmann R. and Flache, A., Understanding complex social dynamics: A plea for cellular automata based modelling. *Journal of Artificial Societies and Social Simulation,* Vol. 1, No3, 1998.

Lazurus, R.S., *Emotion and Adaptation,* Oxford University Press, 1991.

Lindsay, P.H. and D.A. Norman, *Human Information Processing: An Introduction to Psychology,* Academic Press, 1972.

Newell A., *Unified Theories of Cognition,* Harvard University Press, 1990.

Norman, D.A., *Twelve Issues for Cognitive Science.* In: Issues in Cognitive Modeling, A.M. Aitkenhead and J.M. Slack (Editors), LEA Press, 1985.

Oatley, K. and Jenkins, J.M, *Understanding Emotions,* Blackwell, 1996.

Ortony, A., G.L. Clore and A. Collins, *The Cognitive Structure of Emotions.* Cambridge University Press, 1988.

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

Port R.F. and T. Van Gelder (Editors), *Mind As Motion,* MIT Press, 1995.

Power, M. and T. Dalgleish, *Cognition and Emotion: From Order to Disorder,* LEA Press, 1997.

Rolls, E.T., *The Brain and Emotion,* Oxford University Press, 1999.

Schneck, D., Music in human adaptation, CogNet Hot Science, http://cognet.mit.edu/, 2000.

Simon, H.A. Motivational and emotional controls of cognition, *Models of Thought,* Yale University Press, 1979.

Sloman, A. and M. Croucher, Why robots will have emotions. *Proceedings of IJCAI7,* 197-202, 1987.

Sloman, A., Motives, mechanisms and emotions, *Cognition and Emotion* 1, 1987.

Sloman, A. Architectural requirements for human-like agents both natural and artificial, In *Human Cognition and Social Agent Technology,* K. Dautenhahn, Benjamins Publishing, 1999.

Solomon, R.C., *The Passions,* Hackett, 1993.

Spaulding, W.D. (Editor), *Integrative Views of Motivation, Cognition and Emotion,* University of Nebraska Press, 1994.

Velásquez, J.D., When robots weep: emotional memories and decision-making, *Proceedings of AAAI-98,* 1998.

Wollheim, R., *On The Emotions,* Yale University Press, 1999.

Wright, I.P., Sloman, A. and Beaudoin, L., Towards a design-based analysis of emotional episodes, *Philosophy Psychiatry and Psychology, Volume 3,* 1996.

Towards Implementing Free-Will

Bruce Edmonds
Centre for Policy Modelling,
Manchester Metropolitan University,
Aytoun Building, Aytoun Street, Manchester, M1 3GH, UK.
http://www.cpm.mmu.ac.uk/~bruce

"Anyone who considers arithmetic methods of producing random digits is, of course, in a state of sin" John von Neuman[1]

Abstract

Some practical criteria for free-will are suggested where free-will is a matter of degree. It is argued that these are more appropriate than some extremely idealised conceptions. Thus although the paper takes lessons from philosophy it avoids idealistic approaches as irrelevant. A mechanism for allowing an agent to meet these criteria is suggested: that of facilitating the gradual emergence of free-will in the brain via an internal evolutionary process. This meets the requirement that not only must the choice of action be free but also choice in the method of choice, and choice in the method of choice of the method of choice etc. This is directly analogous to the emergence of life from non-life. Such an emergence of indeterminism with respect to the conditions of the agent fits well with the 'Machiavellian Intelligence Hypothesis' which posits that our intelligence evolved (at least partially) to enable us to deal with social complexity and modelling 'arms races'. There is a clear evolutionary advantage in being internally coherent in seeking to fulfil ones goals and unpredictable by ones peers. To fully achieve this vision several other aspects of cognition are necessary: open-ended strategy development; the meta-evolution of the evolutionary process; the facility to anticipate the results of strategies; and the situating of this process in a society of competitive peers. Finally the requirement that reports of the deliberations that lead to actions need to be socially acceptable leads to the suggestion that the language that the strategies are developed within be subject to a normative process in parallel with the development of free-will. An appendix outlines a philosophical position in support of my position.

1 Introduction

To paraphrase the von Neuman quote above: *anyone who considers computational methods of implementing free-will is, of course, in a state of sin.* By simply suggesting that free-will could be implemented I will already have offended the intellectual sensibilities of several groups of people: I will have offended "hard" determinists by suggesting that free-will is possible; I will have offended those who think that free-will is a uniquely human characteristic; and I will have offended those who see free-will as something that is simply beyond design. I have some sympathy for the later two groups – at the moment a human being is the only system that clearly exhibits this facility; and, as will be explained, I do think that free-will can not be *directly* designed into an entity.

Despite almost everybody agreeing that it is fundamentally impossible, arithmetic methods of producing random numbers have become, by far, the most widely used method. These methods (used correctly) are efficient and reliable. We rely on their effective randomness in many cryptographic techniques which, in turn, are relied upon in electronic commerce and the like. Maybe it is time to let the evidence take precedence over assumptive theory – if theory disagrees with practical evidence it is the theory that should change. What was assumed to be a state of sin can turn out to be inspired.

In this paper I will outline a practical architecture that, I argue, could result in a computational entity with free-will. I will start by rejecting extremely idealised conceptions of free-will and suggest instead a more practical set of properties. Then, in section 3, I will put forward the central idea of the paper which is to allow free-will to evolve in a brain during its lifetime. The following 4 sections (4, 5, 6 and 7) consider other

[1] Reportedly said by John von Neuman in 1951 at a conference on Monte Carlo methods.

necessary aspects of the architecture: open-ended development; the co-evolution of strategies against competitive peers; the meta-evolution of the evolutionary process itself; and the necessity of being able to anticipate the consequences of candidate actions. Section 8 then looks at some societal aspects that might allow the development of a framework of acceptable rationality within which free-will can operate. I summarise the suggested architecture in section 9 and finally conclude in section 10. For those who feel philosophically short-changed by this paper there is an appendix which briefly outlines my philosophical position.

2 Conceptions of free-will

It is inevitable that in any implementation process one will move from an *idealised* to a *realised* conception of what one implements. Thus here I am not so interested with artificial or idealised conceptions of free-will, determinism, randomness etc. but with more practical concerns. For if a certain conception of free-will makes no *practical* difference then it is irrelevant to a discussion about implementation (and quite possibly to everything else as well). For if it is impossible to tell whether an entity has a certain property and that entity can do all the things without that property as with it, how can it be relevant in practice?

From this practical perspective, free-will is something that a normal adult human has but an newly fertilised human embryo hasn't. It means that an agent is free to choose according to its will, that is to say that sometimes it is its deliberations on how to achieve its goals that determine its actions and not just its circumstances (including past circumstances).

Of course, many aspects of traditional philosophical analyses of free-will are relevant *if* one avoids the pitfalls of extreme idealisation. For example the points listed below come from philosophy, but are formulated with practical concerns in mind:

(A) The process of deliberation leading to a choice of action has to be free in the sense that it is not constrained to a particular "script" - this means that there is also some *choice* in that deliberation, as well as choice in how to make that choice, and choice in how to make the choice in how to make that choice etc.;

(B) In some circumstances, if others with whom the entity is competing are able to effectively predict its actions they may well exploit this in order to constrain its choice to its detriment - thus it can be important that actions are not predictable by others;

(C) In order for an entity's will to be effective it has to be able to perform some processing that tends to result in actions that (as far as it can tell) furthers its goals – in particular it needs to be able to consider the likely consequences of different possible strategies and *choose* amongst them with a view to furthering its goals;

(D) It must be possible that sometimes it might have taken a different action to those actually taken - that is, given *indistinguishable* circumstances, it would not simply repeat past decisions (even if it did not recall them).

(E) In order to have an entity's decisions allowed by a society of peers it is often necessary that it is able to give an account of its reasons for actions that impinge upon that society, reasons that would be deemed acceptably rational - for those that are not reliably rational can pose a danger to the rest and hence may be prevented from taking certain actions.

These are the criteria I will take to guide my thoughts about implementation rather than abstract issues of theoretical determination and the like. They seem to capture the more *important* aspects of free-will – the aspects of free-will that make it worth having (Dennett 1984).

This is a similar approach to that of Aaron Sloman's (1992), except that it focuses more upon a single issue: ***how can we develop an agent whose decisions are determined by its deliberations and not completely constrained by its circumstances.*** He is right to point out that an entity's decisions can be constrained in different ways and is dependent upon the capabilities and structure of the entity. However the multiplicity of factors does not dissolve the central issue which is concrete and testable; for any entity placed in the same circumstances one can *measure* the extent to which entity acts in the same way[2] and (with humans) collect indirect evidence (by interview) to see the extent to which the actions correlated with the prior deliberations.

3 Evolving free-will in a brain

The basic idea I am proposing, is to provide, in a constructed brain, an environment which is conducive to the *evolution* of free-will in that brain. In this evolutionary process practical indeterminacy emerges first in infinitesimal amounts and then develops into full-blown adult free-will by degrees. This evolution happens in parallel to the development of rationality in

[2] A practical way of putting this is: *as a circumstance approaches a previous circumstance in similarity, does the probability of the a different action resulting decrease to zero?*

the individuality, so that the result is a will which is internally coherent in furthering its goals but yet not determined by its circumstances.

Those who insist that free-will requires prior free-will (arguing that otherwise the choice process would also be determined) can follow the chain of causation (and indeterminism) backwards until it slowly diminishes down a limit of nothing (determinism). In this model the gradual emergence of free-will in the brain is analogous to the emergence of life - it can start from infinitesimal amounts and evolve up from there. This requires that free-will can come in different *degrees* – that circumstances can constrain behaviour to different extents from totally (determinism) to partially (some degree of indetermination). The artificiality of an all-or-nothing division into *having it* or *not* makes as little sense with free-will as it does with life, especially if one is discussing mechanisms for its appearance (as must occur somewhere between the newly fertilised embryo and the adult human. As Douglas Hofstadter said (1985):

> *Perhaps the problem is the seeming need that people have of making black-and-white cutoffs when it comes to certain mysterious phenomena, such as life and consciousness. People seem to want there to be an absolute threshold between the living and the nonliving, and between the thinking and the "merely mechanical,"...*

Thus a situation where free-will evolves in increasing effectiveness during the development of the brain satisfies the first of my criteria. Not only can the actions be free, but also the deliberation that resulted in those actions be free and the process to develop those deliberations be free etc. The fact that the chain of free-will disappears back into the internal evolutionary process can be expressed as a closure property.

The selective advantage that this feature confers upon us (as a species) is primarily that of external unpredictability (combined with an internal rationality). That is in a competitive environment, if an opponent can predict what you will do then that opponent would have a distinct advantage over you. Such competition in a social setting has been posited as one of the evolutionary selective factors that promoted intelligence in our species (Byrne and Whiten, 1988).

That unpredictability *can* be evolved has been shown by Jannink (1994). He developed a simulation with two separate populations which were co-evolved. The first of these populations was allocated fitness on the basis of the extent to which its programs successfully predicted the output of programs from the other and individuals from the second were allocated fitness to the extent that it avoided being predicted by individuals from the first population. Here the two populations are involved in a basic evolutionary 'arms-race'.

Thus the basic architecture I am suggesting is composed of the following elements:

- A framework for decision making processes;
- A population of processes within this framework;
- A way to construct new processes as a result of the action of existing decision making processes and the knowledge of the agent;
- A selection mechanism that acts to (1) select for those processes that tend to further the individual's goals and (2) to select against those processes that are predictable by others.

This evolutionary architecture is the basis for the suggested implementation. However, this architecture needs several more features in order to realise its potential. These are now discussed.

4 Open-ended strategy evolution

In a standard Genetic Algorithm (GA) following Holland (1975), the genome is a fixed length string composed of symbols taken from a finite alphabet. Such a genome can encode only a finite number of strategies. This finiteness imposes a ceiling upon the possible elaboration of strategy. This can be important where individuals are involved in the sort of modelling "arms-race" that can occur in situations of social competition, where the whole panapoly of social manouveurs is possible: alliances, bluff, double-crossing, lies, flattery etc. The presence of a complexity ceiling in such a situation (as would happen with a GA) can change the outcomes in a qualitatively significant way, for example by allowing the existence of a unique optimal strategy that can be discovered.

This sort of ceiling can be avoided using an open-ended genome structure as happens in Genetic Programming (GP) or messy genetic algorithms. Within these frameworks, strategies can be indefinitely elaborated so that is it possible that any particular strategy can be bettered with sufficient ingenuity. Here I use the GP paradigm, since it provides a sufficiently flexible framework for the purpose in hand. It is based upon a tree-structure which is expressive enough to encode almost any structure including neural-networks, Turing complete finite automata, and computer programs (Spector et al. 1999).

The GP paradigm means that the space of possible strategies is limited only by computational resources. It also has other properties which make it suitable for my purposes:

- The process is a path-dependent one since the development of new strategies depends upon the resource of present strategies, providing a continuity of development. This means that not only can completely different *styles* of strategy be developed but also different ways of approaching (expressing) strategies with similar outcomes.

- The population provides an implicit sensitivity to the *context* of action – different strategies will 'surface' at different times as their internal fitnesses change with the entities circumstances. They will probably remain in the population for a while even when they are not the fittest, so that they can 're-emerge' when they become appropriate again. Thus agents using a GP-based decision-making algorithm can appear to 'flip' rapidly between strategies as circumstances make this appropriate.

5 Meta-evolution

Such a set-up does mean that the strategy that is selected by an agent is very unpredictable; what the currently selected strategy is can depend upon the history of the whole population of strategies due to the result of crossover in shuffling sections of the strategies around and the contingency of the evaluation of strategies depending upon the past circumstances of the agent. However the *method* by which new strategies are produced is not dependent upon the past populations of strategies, so there is no backward recursion of the choice property whereby the presence of free choice at one stage can be 'amplified' in the next.

Thus the next stage is to include the operators of variation in the evolutionary process. In the Koza's original GP algorithm there are only two operators: propagation and tree-crossover. Instead of these two operators I suggest that the population of operators themselves are specified as trees following (Edmonds 1998). These operators are computationally interpreted so they *act upon* strategies in the base population to produce new variations. The operators are allocated fitness indirectly from the fitnesses of the strategies they produce using the "bucket-brigade" algorithm of Holland (1975) or similar (such as that in Baum 1998, which is better motivated).

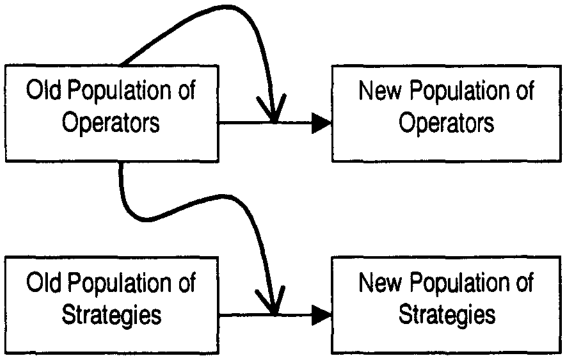

Old Population of Operators
New Population of Operators
Old Population of Strategies
New Population of Strategies

Figure 1. One step of a reflective meta-genetic process

To complete the architecture we set the population of operators to also operate on themselves in order to drive the production of new operators. Now the decision making processes (including the processes to produce the processes etc.) are generated internally, in response to the twin evolutionary pressures of deciding what to do to further the agents goals (in this case profit) and avoiding being predictable to other agents. This is illustrated in Figure 1.

6 Anticipatory rationality

If an agent is to reflectively *choose* its action rather than merely *react* to events, then this agent needs to be able to *anticipate* the result of its actions. This, in turn, requires some model of the world, i.e. some representation of the consequences of actions that has been learnt through past interaction with that world (either via evolution of the entity).

The models of the consequences of action are necessarily separate from the strategies (or plans) for action. It is possible to conflate these in simple cases of decision making but if an entity is to *choose between* plans of action with respect to the expected outcome then this is not possible. There is something about *rationality* which excludes the meta-strategy of altering one's model of the world to suit ones chosen strategy - the models are chosen according to their accuracy and relevance and the strategies are *then* chosen according to which would produce the best anticipated outcome according to the previously selected world model.

A reactive agent may merely work on the presumption that the strategies that have worked best in the past are the ones to use again. This excludes the possibility of anticipating change or of attempting to deliberately 'break-out' of current trends and patterns of behaviour.

Thus we have a process which models the consequences of action and one which models strategies for action. To decide upon an action the best relevant model of action consequence is chosen and the various strategies for action considered with respect to what their anticipated consequences would be if the consequence model is correct. The strategy that would seem to lead to the consequence that best fitted the goals would be chosen. This is illustrated in figure 2 below.

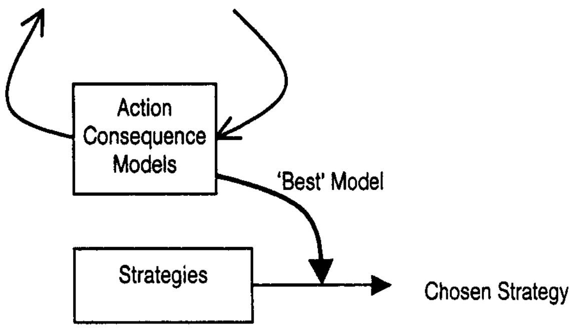


Figure 2. Using anticipation with strategy selection

7 Co-evolution

The next important step is to situate the above elaborated model of strategy development in a society of competitive peers. The development of free-will only makes sense in such a setting, for if there are not other active entities who might be predicting your action there would be no need for anything other than a reactive cognition. This observations fits in with the hypothesis that our cognitive faculties evolved in our species due to a selective pressure of social origin (Byrne and Whitten, 1988).

Thus we have a situation where many agents are each evolving their models of their world (including of each other) as well as their strategies. The language that these strategies are limited to must be sufficiently expressive so that it includes strategies such as: attempting to predict another's action and doing the opposite; evaluating the success of other agents and copying the actions of the one that did best; and detecting when another agent is copying one's own actions and using this fact to do what would help you. Thus the language has to have 'hooks' that refer to ones own actions as well as to other's past actions and their results.

In circumstances such as these it has been observed that agents can spontaneously differentiate themselves by *specialising* in different styles of strategies (Edmonds, 1999). It is also not the case that just because these agents are competing that they ignore each other. Such a co-evolution of strategy (when open-ended and resource limited) can result in the intensive use of the actions of others[3] as inputs to their own deliberation, but in a way that is unpredictable to the others (Edmonds, in press). So that the suggested structure for agent free-will can include a high level of social embedding.

8 Structuring the development of free-will within a society of peers

The final difficulty is to find how to structure this mental evolution so that in addition to maintaining the internal coherence of the deliberations and their effectiveness at pursuing goals and being unpredictable to others, the actions of the agent can be presented to others as rational and verified as such by those agents. This is in order to fulfil criterion (E) above.

This last criterion can be achieved if there is a normative process which specifies a *framework* of rationality which is not restrictive so that different deliberative processes for the same action can be simultaneously acceptable. The framework must be loose enough so that the *openness* of the strategy development process is maintained, allowing creativity in the development of strategies, etc. But on the other hand must be restrictive enough so that others can understand and empathise with the deliberative processes (or at least a credible reconstruction of the processes) that lead to action.

There are number of ways in which this framework could be implemented. I favour the possibility that it is the *language* of the strategies which is developed normatively in parallel with the development of an independent free-will. Thus the *bias* of the strategies can be co-evolved with the biases of others and the strategies developed within this bias.

9 Putting it all together

Collecting all these elements together we have the following parts:

- A framework for the expression of strategies which is (at least partially) normatively specified by the society of the entity.
- An internal open-ended evolutionary process for the development of strategies under the twin selective pressures of favouring those that further the goals of the entity and against those that result in actions predictable by its peers.
- That the operators of the evolutionary process are co-evolved along with the population of strategies so that indeterminism in the choice of the entity is amplified in succeeding choices.
- That models of the consequences of action be learned in parallel so that the consequences of candidate strategies can be evaluated for their anticipated effect with respect to the agent's goals.

Each of these elements have been implemented in separate systems, all that it requires is that these be put

[3] Including the communicative acts of others

together. No doubt doing this will reveal further issues to be resolved and problems to be solved, however doing so will represent, I suggest, real progress towards the goal of implementing free-will.

10 Conclusion

Although it is probably not possible to implement the facility for free-will directly in an agent (i.e. by *designing* the detail of the decision making process)[4], I have argued that it is possible to implement a cognitive framework within which free-will can *evolve*. This seems to require certain machinery: an open-ended evolutionary process; selection against predictability; separate learning of the consequences of action; anticipation of the results of action and the evolution of the evolutionary process itself. Each of these have been implemented in different systems but not, as far as I know, together.

The free-will that results is a practical free-will. I contend that if the architecture described was implemented the resulting facility would have the essential properties of *our* free-will from the point of view of an external observer. Such a facility seems more real to me than many of the versions of free-will discussed in the philosophical literature, because it is driven more by practical concerns and observations of choice and is less driven by an unobtainable wish for *universal* coherency.

There are basically three possibilities: free-will is a sort of 'magic'; it is an illusion; or it is implementable. I hope to have made the third a little more real.

Acknowledgements

Some of the cognitive models mentioned in this paper were implemented in the modelling language SDML[5]. SDML has been developed in VisualWorks 2.5.1, the Smalltalk-80 environment produced by ObjectShare. Free distribution of SDML for use in academic research is made possible by the sponsorship of ObjectShare (UK) Ltd. The research reported here was funded by the Faculty of Management and Business, Manchester Metropolitan University.

Philosophical appendix

There is no stopping some people philosophising, however inappropriate or unhelpful this is in particular contexts. Such people seem to think that it is both possible and useful to formulate generally and reliably true principles (i.e. principles completely without exceptions regardless of context) about the world using argument. For these people I *briefly* outline my position below, for full details they will have to come and argue with me.

- The "hard" deterministic thesis is untestable and has no practical consequences – the world is equally explained using it or otherwise since we can not rewind the world to see whether this thesis does in fact hold. The only consequences it can thus have is to sanction a normative claim about the use of the term "free-will" – this amounts to no more than a position that *given I conceive of the world as determined then there can not be anything denoted as an indeterministic process including free-will.*
- The above point can be demonstrated by considering the following thought-experiment: compare a human who had a 'real indeterminism pump' with an otherwise identical human with only a good 'pseudo-random' generator – there would be no testable or practical difference between them. The distinction is thus irrelevant except in how we conceive of our world.
- There is a lot of evidence against the hard deterministic thesis both at the micro-level (quantum effects) and at the macro level (that many complex systems are not determinable in practice).
- Any strengthening of the deterministic thesis to make it actually applicable (e.g. that given *almost* identical circumstances a certain identifiable system will exhibit the same behaviour) renders it false when applied to some systems – for example humans will not always exhibit the same behaviour in practically indistinguishable circumstances (even if they do not recall their previous decisions).
- I can see no reason why an *indeterministic* process has to be *arbitrary*[6].
- It is difficult to see how any conception of free-will that did not come down to the principles (A)-(E) above could have any realisable meaning.
- It is very difficult to see how the facility of free-will evolved in us as a species if it was not implementable and was inextricably linked with its practical consequences so it could be selected for.

[4] Indeed in another paper to be presented at the "Starting from Society symposium here at AISB2000, I argue that *intelligence* (as defined by the Turing Test) cannot be designed according to an explicit plan.

[5] Strictly Declarative Modelling Language - see http://www.cpm.mmu.ac.uk/sdml/

[6] although it possible that it is to all practical purposes *random* from the point of view of an external observer, depending upon the conception of randomness.

- It is much more useful (in the analysis of issues surrounding free-will) to consider the practical sources, advantages and consequences of different *kinds* of free-will as argued in Dennett (1984) and Sloman (1992).

Thus the practical, common-sense conception is a better representation of free-will than many of the idealised, philosophical characterisations of it. When involved in a process of implementation, it is wise to base one's work on the *best* representation available.

References

Baum, E. Manifesto for an Evolutionary Economics of Intelligence. In Bishop, C. M. (ed.), *Neural Networks and Machine Learning*, Springer-Verlag, 285-344, 1998.

Byrne, R. W. and Whiten, A. (eds.) Machiavellian Intelligence: social expertise and the evolution of intellect in monkeys, apes, and humans, Oxford: Clarendon Press, 1988.

Dennett, D. C. Elbow Room: varieties of free-will worth having. Oxford: OUP, 1984.

Edmonds, B. Meta-Genetic Programming: Co-evolving the Operators of Variation. CPM Report 98-32, <http://ww.cpm.mmu.ac.uk/cpmrep32.html>, 1998.

Edmonds, B. Gossip, Sexual Recombination and the El Farol bar: modelling the emergence of heterogeneity. *Journal of Artificial Societies and Social Simulation*, **2**(3), <http://www.soc.surrey.ac.uk/JASSS/2/3/2.html>, 1999.

Edmonds, B. Capturing Social Embeddedness: a constructivist approach. *Artificial Behavior*, **7**(3/4), in press.

Harnad, S. Turing Indistinguishability and the Blind Watchmaker. In: Mulhauser, G. (ed.) *Evolving Consciousness*, Amsterdam: John Benjamins, in press.

Hofstadter, D. R. Analogies and Roles in Human and Machine Thinking, In *Metamagical Themas*, New York: Basic Books, 1985.

Holland J. H. Adaptation in Natural and Artificial Systems : an introductory analysis with applications to biology, control, and artificial intelligence. Ann Arbor : University of Michigan Press, 1975.

Jannink, J. Cracking and Co-evolving Randomizers. In Kinnear, K. E. (ed.) *Advances in Genetic Programming*, Cambridge, MA: MIT Press, 425–444, 1994.

Koza, J. R. Genetic Programming: on the programming of computers by means of natural selection, Cambridge, MA: MIT Press, 1992.

Palmer, R. G., Arthur, W. B., Holland J. H., et al. Artificial Economic Life: a simple model of a stockmarket, *Physica D* **75**:264-274, 1994

Sloman, A. How to Dispose of the Free-Will Issue. *AISB Quarterly*, **82**, Winter 1992-3, 31-32, 1992.

Spector, L., Langdon, W. B., O'Reilly, U-M., and Angeline, P. J. (eds.) *Advances in Genetic Programming, Volume 3*, Cambridge, MA:MIT Press, 1999.

Making a mind: a cognitive engineering approach

John Fox
Advanced Computation Lab
Imperial Cancer Research Fund
Lincoln's Inn Fields
London UK
jf@acl.icnet.uk

Abstract

In recent years empirical research into AI models of mind have has fallen off in favour of formal and engineering studies of intelligent systems. We argue the need for more "scientific" investigations of mind, with a strong empirical as well as theoretical methodology, and advocate a unified research programme combining knowledge level and architectural theories.

1 Introduction

Before the Japanese 5th, Generation Computer Project stimulated an explosion of interest in the 1980s the AI research community was quite small and focused. Although the field was clearly interdisciplinary, the computer scientists, psychologists, philosophers and others involved had enough unity of purpose to agree to call themselves "cognitive scientists", sharing a common interest in understanding the nature of intelligence, both natural and artificial.

A direct effect of the 5th Generation Project was to draw researchers into AI from many fields that did not have a traditional interest in cognitive science. These included statisticians, operations researchers and software (and other kinds of) engineers. They brought many different perspectives and techniques to the field, but a consequence was that it became fragmented into competing schools of thought. Debates raged, including those of symbolicists vs connectionists, logicists vs Bayesians and experimentalists vs formalists. Indeed a fundamental philosophical split also emerged, between those communities who emphasised experimental *science* (e.g. psychology and neuro-science), *practical engineering* (e.g. knowledge systems and robotics) and *formal theory* (e.g. mathematical logic and bayesian inference). Although one may argue that the diversity of approach was healthy it seems to have had some negative effects, and in particular some loss of scientific consensus about what the field of AI was about.

For reasons which are unclear, but we hope because of a weariness with unproductive factionalism, AI may now be trying to recover the scientific unity which it lost in the nineteen eighties. Proposals for "unified theories of cognition", "intelligent agents", "theories of consciousness" – and this conference – suggest a desire to return to the intellectual roots of the subject; to understand the principles of mind(s).

We are experimental scientists (though with strong interests in practical engineering and formal theories in Al) so we have always had a strong interest in empirical approaches to cognitive science. We have had a particular interest in the experimental and theoretical programme advocated by the late Allen Newell, which was aimed at constructing a "unified theory of cognition" (UTC; Newell, 1990). Frustrated with psychology's tendency to fragment into niche areas like memory, reasoning, decision-making, perception, language and so forth, Newell argued that the time is right to exploit computational ideas to bring these together in understanding complete systems. Terms like "mind" were unfashionable at the time but understanding and building minds was certainly what Newell was about.

2 Understanding and building minds: the SOAR programme

Newell's concept of a unified theory is to be found in its most developed form in the SOAR theory of the human "information processing architecture". SOAR grew out of a long collaboration on human problem solving with Herbert Simon (Newell and Simon, 1972), which stimulated many conceptual and technical developments in cognitive science. The clearest instance of this was the realisation that a very simple computational element, the humble production rule, was sufficient to implement a powerful and important family of symbol-manipulating systems. SOAR built on the successes of earlier production rule technologies like PSG and OPS, by making a couple of small but important extensions to the elementary recognise-act cycle of rule interpretation. The main extensions were the ability to generate *goals* dynamically and to *learn* new rules based on experience. Laird, Newell and Rosenbloom (1987) argued that the resulting computational architecture had the necessary and sufficient means to demonstrate "general intelligence". SOAR was presented as an architecture for general intelligence and later as a general tool for building "expert" and other knowledge systems. The SOAR team set about a research programme to demonstrate this, both as an effort in cognitive science and in cognitive engineering.

Despite our interest in unified theories of cognition and our sympathy for the goals of the SOAR research programme we think that much of the work that followed was conceptually and methodologically problematic. SOAR played an important part in establishing rule-based programming as a major AI paradigm (along with work on theorem provers, logic programming languages like Prolog and expert system shells), but as a contribution to the science of mind and as a platform for building artificial intelligences it is open to criticism.

SOAR was certainly an admirable attempt to capture many of the manifestations of mind with a simple computational architecture. For a while it was also a credible theory of the human cognitive system. However, this credibility was quickly eroded, particularly among psychologists who were more concerned with empirical facts about the human mind than, say, computational effectiveness or parsimony. Although Newell himself took the problem of explaining empirical observations seriously the evidence base for SOAR as a theory of the human mind was limited, and even somewhat selective, as discussed by Cooper and Shallice (1995).

In fact, from our perspective as traditional "natural scientists" the SOAR programme did not look like good science at all. Although Newell speaks of SOAR as an incremental, "Lakatosian" programme of research into unified theories he proposed no methodology for systematic development and testing of such theories. Indeed SOAR itself became something of an ideology (surely the antithesis of science) with its central tenets often justified on "sufficiency" criteria rather than empirical or theoretical grounds, and strongly defended against revision.

Following Allen Newell's death there seems to have been repositioning of SOAR, which is now less a unified theory of cognition and more a platform for engineering knowledge systems and applications. Even as an engineering tool, however, we wonder if SOAR has lost its way. As the field of knowledge engineering has moved on to develop powerful high-level tools for AI applications, like deductive databases, object-oriented knowledge bases and agents. SOAR has remained a "pure" production rule technology. Given Newell's emphasis on the importance of the "knowledge level" for cognitive science it is also surprising that successful knowledge engineering techniques have not been incorporated into SOAR. We are thinking here of techniques like generic tasks and ontologies, and formal specification and verification of knowledge bases (see Fox and Das, 2000, for a review).

In the end of course the UTC community might reasonably argue that SOAR's significance is that it shows how you can build a mind from many little parts" to quote Marvin Minsky. It embodies an "architecture for general intelligence" constructed from the most minimal elements: goal-based problem-solving, simple if-then rules, a recognise-act rule execution cycle with conflict resolution, and learning by chunking new rules as problems are solved and goals achieved. Unfortunately even this claim is ambiguous, since none of these mechanisms is demonstrably *necessary* for general intelligence. Furthermore implementations of SOAR were generally (very) large LISP or C programs, in which the basic computational principles of the theory were mired in implementation details whose role and importance for its capabilities were hard to assess (Cooper, Fox, Farringdon and Shallice, 1996).

To summarise, we are not convinced that the research methods that have been used by the SOAR community are sufficient to provide a firm scientific basis for understanding naturally occurring minds, or a rigorous engineering methodology for building artificial ones. However we do believe that the goals of the SOAR community were important and we don't wish to pick on SOAR as uniquely unsatisfactory in these respects. In fact the AI community generally has had a tendency to fall between the stools of good empirical and theoretical science and good practical engineering. However, if we are truly moving into a new period of research into minds and other kinds of unified theories then we need to adopt more systematic and articulated research methods than we have in the past.

3 Empirical programmes for understanding minds

Despite the doubts about SOAR we continue to share Newell's general aspiration to develop a unified theory of cognition, and his conviction that symbol processing is sufficient (and in some respects necessary) to understand general intelligences, and the human mind specifically. Our own work has been aimed at similar goals with rather different methods. This has emphasised systematic programmes of investigation into intelligent, knowledge-rich behaviour, seeking insights from how intelligence is manifested and constrained in complex, real-world settings.

If we can be allowed a little post *hoc* rationalisation we have adopted a programmatic approache, which can be summarized as a number of steps:

1. Define a range of cognitive functions that together are taken to embody "mind" (the theoretical *scope)*
2. Select a subset of these functions for study (a tractable *focus* for the investigation)
3. Adopt a field of knowledge (as a *domain* for the research programme)
4. Carry out objective studies of behaviour in the domain (keep the programme honest)
5. Develop a computational theory for the selected focus
6. Assess the theory against objective data and appropriate evaluation criteria

We have pursued two complementary research programmes based on this slightly sanitized model. One has been concerned with human reasoning and decision making (mostly in medicine), and one has been aimed at developing practical tools for supporting reasoning, decision-making and action (also in medicine). The motivation for this two-pronged strategy is that "mind" is a complex, multi-faceted concept that probably cannot (as yet) be understood within any one explanatory framework. Consequently a full understanding can only be achieved by exploring it with multiple theoretical tools and concepts.

Three established frameworks for understanding minds that have been developed in cognitive science are illustrated in figure 1. Our own work is attempting to address the unification of problem-solving, reasoning, decision-making, planning and other cognitive functions from epistemic, functional and structural perspectives, though we shall only discuss two programmes in this paper. These are concerned with *architectural* issues and the application of *knowledge* by different architectures.

The first research programme is addressed to the *knowledge level* seen at top of figure 1. (We prefer the term "epistemic", as this subsumes ontological and cultural as well as individual knowledge.) This has been specifically considered in the domain of medicine but the emphasis has been on building computational models of realistically complex expertise that can be put to practical use in any domain. The programme instantiates the general schema above as follows:

1. Theoretical scope – perception, memory, knowledge representation, reasoning, problem-solving, decision making, planning, action control, and learning
2. Focal subset – reasoning, decision-making, scheduling, acting
3. Knowledge domain – patient care in cancer and other complex medical domains
4. Observational studies – development and evaluation of medical decision aids
5. Theory – logical models of knowledge and expertise
6. Evaluation – effectiveness of decision aids on real world (medical) problems

The second programme follows the SOAR community and others in psychology in trying to understand the *human information processing* architecture (lower level in figure 1), focussing on how people learn to carry out reasoning and decision-making in tasks similar to medical diagnosis. The programme instantiates the general research schema as follows.

1. Theoretical scope (as before) – perception, memory, knowledge-representation, reasoning,

problem-solving, decision making, planning, action control, and learning

2. Focal subset – memory, reasoning, decision-making, action control, learning
3. Knowledge domain – medical diagnosis
4. Observational studies – controlled experiments of human subjects carrying out a simulated diagnosis task
5. Theory: computational model of human information processing architecture
6. Evaluation: comparison of behaviour of human subjects and model

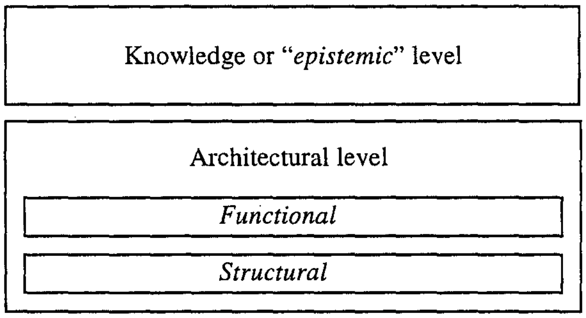


Figure 1: Explanatory levels of cognitive processes

We distinguish here between functional level of description (dealing with cognitive functions like memory, reasoning, perception and the like) and structural levels of description that are concerned with implementation in brain tissue, silicon technologies or whatever. Our own concern has been primarily with functional views, though minds can of course be investigated by neuroscientists and others who are more interested in their physical embodiments.

4 Results to date

The first programme of research is focussed on a "knowledge level" description of mind. In a sense we have developed a *competence* model of medical expertise without incorporating performance limitations such as people's limited information processing capacity or our tendency to forget things. Types of expertise have ranged from decision making in drug prescribing and risk assessment to the execution of complex clinical procedures such as breast cancer screening and chemotherapy which involve many tasks carried out over time (Fox and Thomson, 1999). As we addressed wider and more complex domains a broad model of expertise emerged, expressed as a generalised epistemic model, that is summarised in figure 2. To clarify this informal model the framework has been formalised using a combination of classical and non-classical logics to define the semantics of the arrows. These represent the kinds of inference by which goals are derived from beliefs, actions from plans etc. (Das et al, 1997; Fox and Das, 2000).

The second research programme is aimed at a functional model of the human information processing architecture (Fox and Cooper, 1997; Cooper refs). In contrast to SOAR, which makes strong commitments to a particular architectural theory, we have used a model-building environment that has been designed to permit the construct and explore the properties of alternative possible computational architectures. This system, called COGENT (Cooper and Fox, 1998) provides a set of standard components for modeling information processing architectures as a set of distinct computational compartments, such as I/O channels, memory buffers, production systems, prolog programs, and connectionist and other learning processes. COGENT provides a range of tools for programming the behaviour and setting parameters of the compartments, and managing communication between them. Such COGENT models are effectively platforms for "running" and exploring the behaviour of an expertise model under different assumptions. While the domino model in figure 2 represents a competence model of human expertise, the COGENT architecture is a performance model that can be used to superimpose features of human cognition on the disembodied epistemlc view.

Figure 3 illustrates an architecture that we have implemented using COGENT, in this case a model of how people acquire and apply knowledge in order to carry out a medical diagnosis task (Cooper et al, 1998). This model is built as a set of basic compartments that implement decision making and learning mechanisms as well as a knowledge base and working memory. Empirical studies of human behaviour on the task allowed us to identify important performance characteristics, such as the tendencies to focus on positive rather than negative information, to be overly influenced by recent information, to forget things, and so forth. The resulting model provided a successful simulation of many aspects of human performance on the diagnosis task.

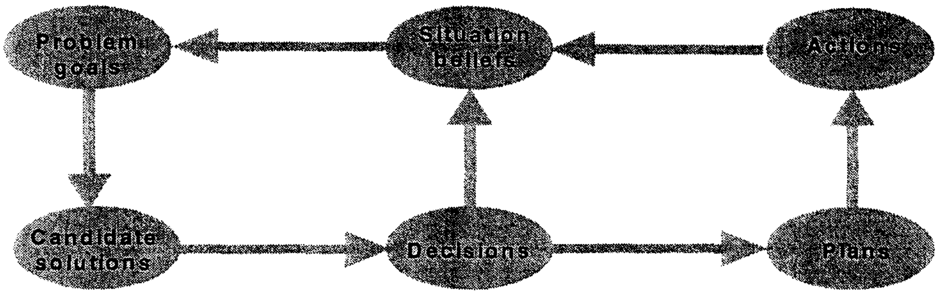


Figure 2: The "domino": a generalised model of reasoning, decision making and planning

The COGENT environment also provides explicit support for systematically exploring alternative theoretical positions. It has proved to be a highly effective tool for comparing models of reasoning and decision making in medical diagnosis, specifically comparing a family of symbolic models of medical judgement similar to the domino model with Bayesian and connectionist models (Cooper Yule and Fox, submitted).

5 Designing a mind: a software-engineering approach

In other work we have shown that symbolic models of reasoning, decision making etc. developed for modeling human information processing and as well as the decision models used in the development of practical decision aids are compatible with the broad framework embodied in the domino model (Fox and Das, 2000). In fact this model seems very general, covering as it does beliefs, goals, decisions, plans, actions and so forth, and we believe therefore that it embodies a simple, but interesting, class of cognitive agents or "minds".

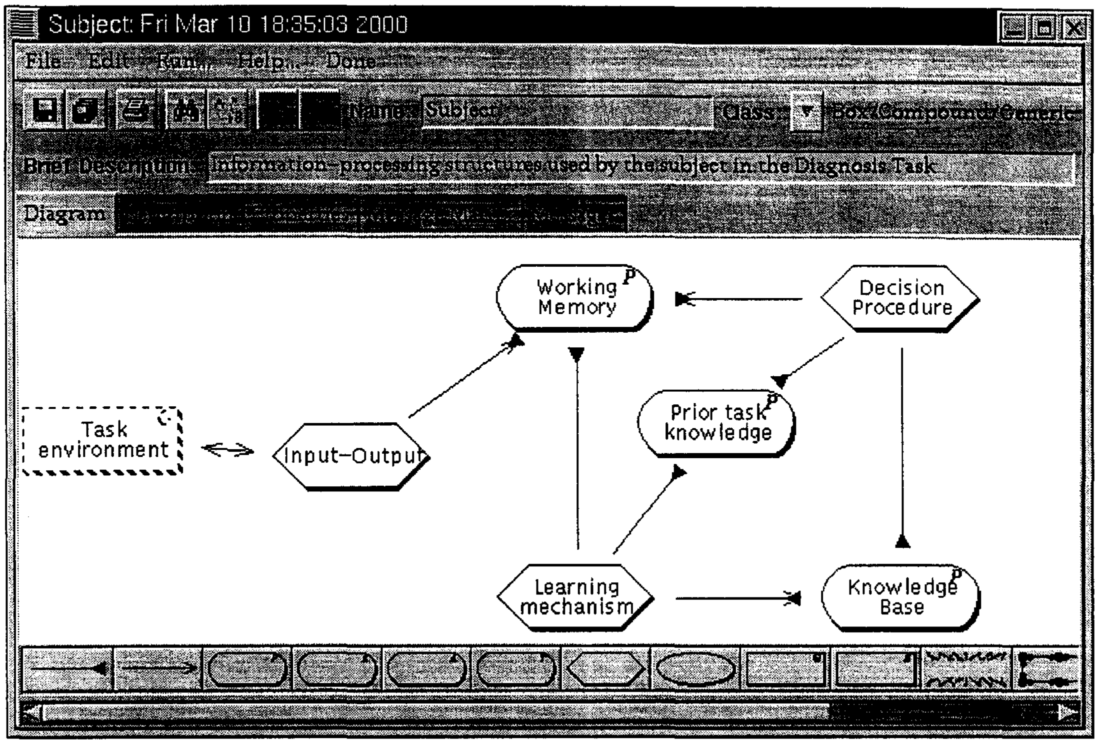
To explore this, we have developed the model further, by embodying it in another development environment whose function is to support the development of expert software agents as distinct from architectures. In order to do this we decided to move from the descriptive cognitive modeling and general AI programming supported by COGENT, to adopt formal design and implementation techniques which we felt would simplify the development of agents and yield insights into their competencies and limitations at the epistemic level. The result, the PRO*forma* development environment (Fox and Das, op *cit)* can be used as a tool for conceiving, designing and implementing cognitive agents based on the domino framework, and potentially for systematically investigating the strengths and weaknesses of such agents.

The domino is primarily concerned with the *logical* processes involved in executing plans, making decisions and scheduling primitive actions. The latter include actions whose goal is to modify the agent's environment *(via* another agent, or a physical device or effector) and actions whose function is to obtain information about the environment (indirectly from another agent or directly via a sensor).

The first stage in turning the model into a practical design tool was to develop a knowledge representation language that embodies the required concepts. In common with other formal knowledge representation languages PRO*forma* is based on first-order logic (van Harmelen et al, 1993) with specific extensions to support constructs like beliefs, goals, decisions and plans. Van Harmelen and Balder (1992) summarize advantages of using such languages for describing the structure and/or behavior of knowledge systems as follows:

- the removal of ambiguity
- facilitation of communication and discussion between designers
- the ability to deduce properties of the *knowledge* independently of a specific *implementation.*

The second stage is to support this language in a development environment that provides a range of computer-aided software engineering tools that support the specification of agent systems in the language. The approach that we have taken to this is to build the environment around a basic ontology of goal-based activities, or tasks (figure 4). This ontology includes four classes of task only: *decisions* (any kind of choice); *actions* and *"enquiries"* which



Subject: Fri Mar 10 18:35:03 2000

Name: Subject

Class: Box/Compound/Generic

Brief Description: Information-processing structures used by the subject in the Diagnosis Task

Figure 3: Information processing architecture for simulating complex cognition

are actions that are intended to modify the agent's environment in some way or return information to an agent, and *plans* (collection or sequence of tasks, including subplans). All tasks share a number of fundamental attributes, where the *goal* of the task, its *preconditions* (which must be true before the task can be executed) and *postconditions* (which will hold if the task is completed successfully) are defined. The general attributes are inherited from the generic or *root* task and particular instances of a task type (e.g. a diagnosis decision, a risk assessment decision) are defined by assigning distinctive values to the attributes. The 4 task subtypes also have distinguishing attributes. For example plans have *components* and *termination* and *abort conditions* while decisions have *candidates* and *argument schemas* for arguing the pros and cons of different candidates (Fox and Das, 2000).

The relationship between the task ontology and the domino model is shown in figure 5. The processes on the left-hand-side of the model represent a generalised process of decision making, by which an agent commits to new beliefs and/or plans, and plans decompose into component plans and actions through a process of task scheduling.

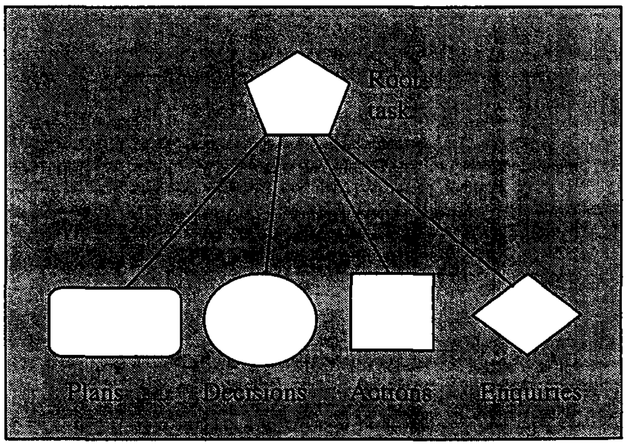


Figure 4: a task ontology for modelling expertise

The PRO*forma* knowledge representation language is a formal, compositional, specification language. It has proved to be an effective formalism for representing complex expertise models. Development of PRO*forma* agents language is supported by a development environment (developed in collaboration with *Infer*Med, Ltd). This incorporates a range of CASE tools for assembling agents from epistemic components and supports a systematic development method (figure 6).

The PRO*forma* language, toolset and development method have been validated on a range of practical applications with good results: the task ontology is small but highly expressive and the tools and method considerably simplifies the job of building epistemic models.

Some of the apparent power of the method comes from the choice of core abstractions (the four task types seem to build naturally into complex cognitive structures) and the use of proven software engineering methods (declarative specification and compositionality). Furthermore, however, the basic theoretical commitments of the language are inspired by psychology and AI (notably the core concepts of symbolic beliefs, goals, arguments, commitments and constraints). From an epistemic position these seem to capture intuitive and widely accepted ideas about important aspects of "mind".

6 Making a mind: towards a systematic research programme

Until now *PROforma* has primarily been used as an engineering tool, to build practical AI systems. However, we believe that we are now in a position to establish a new empirical and theoretical programme to systematically explore the types of agent that can be constructed within such theoretical frameworks. One such programme could draw simultaneously upon PRO*forma* as a way of modeling the epistemic level, and COGENT for the implementation level. The aim of such a programme would be a deeper understanding of the kinds of intelligences that can be implemented with this group of concepts: strengths and weaknesses, capabilities and failure modes etc.

We believe that the architectural and the epistemic programmes have both yielded significant insights into these different aspects of mind and our goal now is to integrate them. To do this we advocate a systematic programme of research in which we systematically vary competence and performance properties and parameters in order to develop a deeper insight into the underlying class of systems of which the human mind is an instance. Among the systematic variations we might carry out are the following.

Competence modelling

1. theoretical scope (e.g. extending generic task sets to include sensor and effector functions)
2. properties of tasks (e.g. symbolic, statistical or connectionist components)
3. compositionality properties (e.g. shared data versus meta-level reasoning over tasks)

Performance modelling

4. execution parameters (e.g. memory capacity, bandwidth of communication channels between compartments)
5. operational reliability (e.g. data losses during communication, memory decay)

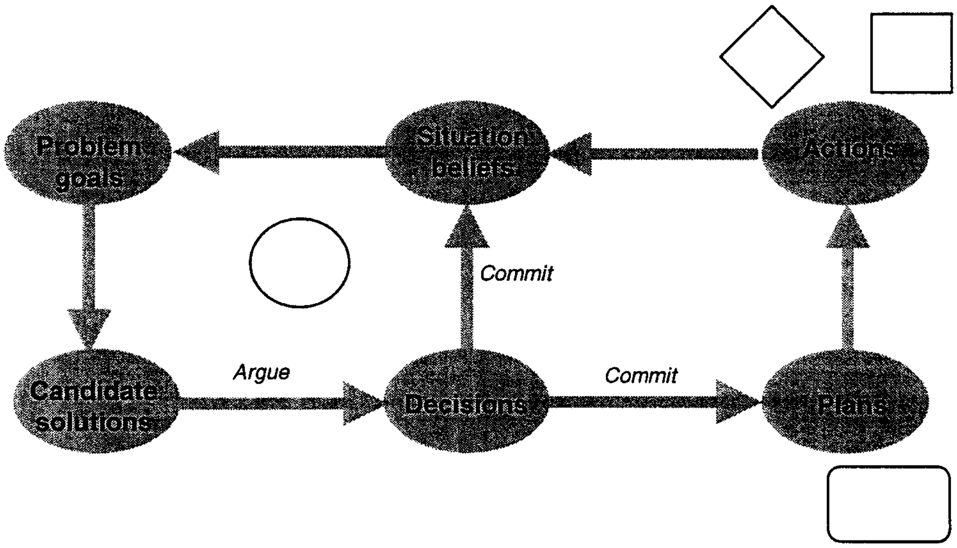
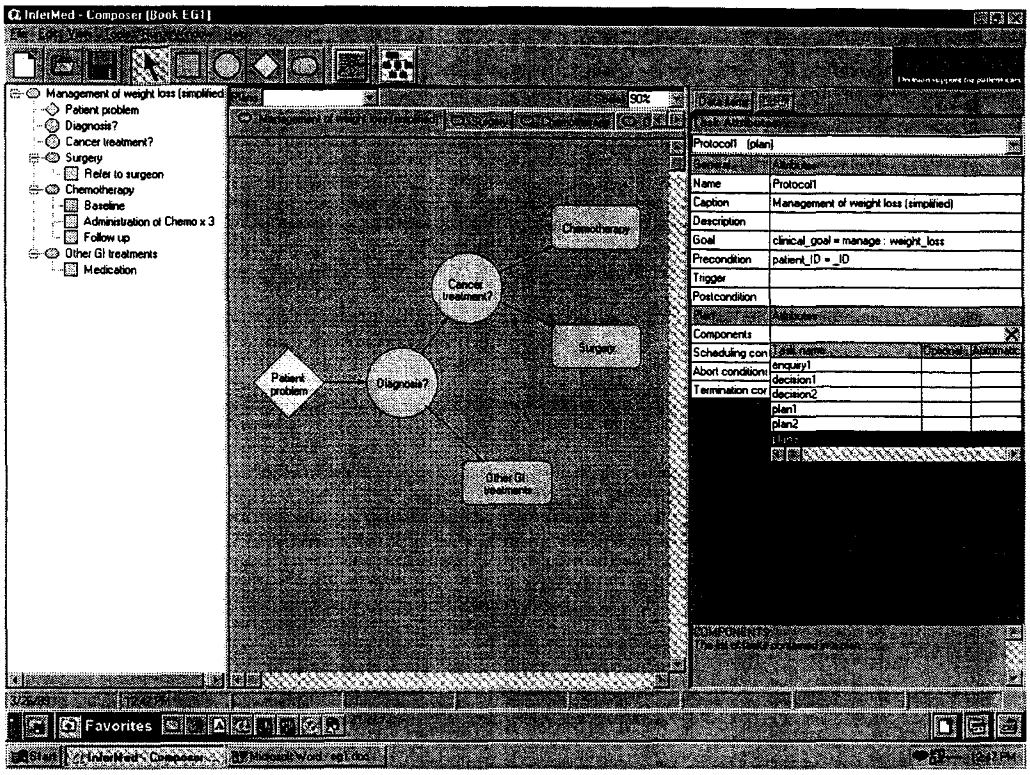


Figure 5: The relationship between the PRO*forma* task set and the functions delineated by the generalized domino model.



InferMed - Composer [Book EG1]

- Management of weight loss (simplified
  - Patient problem
  - Diagnosis?
  - Cancer treatment?
  - Surgery
    - Refer to surgeon
  - Chemotherapy
    - Baseline
    - Administration of Chemo x 3
    - Follow up
  - Other GI treatments
    - Medication

Protocol1 (plan)

Name	Protocol1
Caption	Management of weight loss (simplified)
Description	
Goal	clinical_goal = manage : weight_loss
Precondition	patient_ID = _ID
Trigger	
Postcondition	
Components	
Scheduling con	
Abort condition:	
Termination cor	

Task name	Optional	Automatic
enquiry1		
decision1		
decision2		
plan1		
plan2		

Figure 6: PRO*forma* environment for building intelligent agents based on domino model. This provides a graphical design tool for sketching a task network (left and centre panels) and populating the task definitions (right panel) using standard generic task models.

We believe that a systematic research programme along these lines would provide insights into the general nature and limitations of a significant class of natural and artificial minds.

This raises an interesting question from the point of view of the present symposium on "making a mind", viz: does *PROforma* embody a systematic method for constructing minds, albeit rather simple ones, and does COGENT offer a platform for systematically investigating the abilities and limitations of such minds? On the face of it the general approach is reasonably consistent with the broad and well-established consensus on computational aspects of mind:

Newell and Simon "We confess to a strong premonition that the actual organization of human programs closely resembles the production system organization" *Human problem solving,* 1972 p 803.

Minsky "How can intelligence emerge from non-intelligence? To answer that, we'll show that you can build a mind from many little parts, each mindless by itself." *The Society of Mind,* 1985.

Franklin "Minds tend to be embodied as collections of relatively independent modules, with relatively little communication between them". *Artificial Minds,* 1995.

Furthermore the theoretical focus of what we are doing is largely uncontroversal. As Allen Newell succinctly puts it

> "Problem solving, decision-making and routine action? These are at the center of cognitive behavior. Memory, learning and skill? These equally belong at the center.... Perception and motor behavior? Matters now get interesting, because they are often excluded. Indeed the term cognition emerged in part to indicate the central processes that were ignored by peripheral perception and motor behavior" *Unified theories of cognition,* 1990, p *15.*

So the combination of PRO*forma* and COGENT seems to represent a reasonable view of what is essential to cognition, though more attention to perception and action will be needed if we are to upgrade our terminology to speak of "mind" without being accused of naievete.

Of course there will still be things missing, such as the conative and affective aspects of mind. However, in this respect we must be patient – scientific theories are always incomplete:

> "Most people still believe that no machine could ever be conscious, or feel ambition, jealousy, humor, or have any other mental life-experience. To be sure, we are still far from being able to create machines that do all the things that people do. But this only means that we need better theories of how thinking works." Marvin Minsky, *Society of Mind,* 1995.

Overall we believe that the idea that mind is composed of many little parts which act together to express "intelligence" remains a viable and arguably the most promising approach to achieving unified theories of cognition. Whether these little parts are as small as production rules or as large as, or perhaps larger than, "tasks" like those of the PRO*forma* task ontology there is much to be learned about how such entities may perform in different implementations.

7 Conclusion

For a limited but significant range of mental functions, like reasoning and decision-making, scheduling and planning, we can now give a reasonable account of how a mind like the human mind may operate, and we can build effective automata with comparable capabilities in complex, knowledge rich domains. There is much more to do, but little reason to doubt that the paradigms that are emerging can take us in productive directions.

If AI and cognitive science are to achieve their original objectives, however, we believe that they must establish more systematic empirical and theoretical programmes than have been pursued lately, and we advocate a broad agenda covering both competence and performance theories of mind. Our preferred tools are PRO*forma* and COGENT but like Allen Newell we believe a *programmatic* approach is what is important for the field. As scientists we should avoid commitments to specific concepts or tools and maintain an open mind.

References

Cooper R, Fox J, Farringdon J, Shallice T "A systematic methodology for cognitive modelling" *Artificial Intelligence,* 85, 3–44, 1996.

Cooper R, Fox J "COGENT: A visual design environment for cognitive modelling" *Behaviour Research Methods, Instrumentation and Computers,* 1998, 30 (4), 5 53–564.

Cooper R, Fox J, Learning to make decisions under uncertainty, *Proc. 19th Annual Conference of the Cognitive Science Society,* 1998.

Cooper R and Shallice T "SOAR and the case for unified theories of cognition" *Cognition, 55,* 115–149, 1995,

Cooper R, Yule P and Fox J "Learning to make decisions: a systematic comparison of three traditions. Submitted.

Das S, Fox I, Elsdon D, Hammond P "A flexible architecture for a general intelligent agent" *Journal of Experimental and Theoretical Artificial Intelligence,* 9, 407–440, 1997.

Fox and Cooper "Cognitive processing and knowledge representation in decision making under uncertainty" in R W Scholz and A Zimmer (eds) *Qualitative aspects of decision making,* Langerich: Pabst, 1997.

Fox J, and Das S *Intelligent agents: from cognitive science to cognitive engineering,* AAAI and MIT Press (in press).

Fox J, and Thomson R "Decision Support and Disease Management: A Logic Engineering Approach", IEEE Transactions on Information Technology in Biomedicine, volume 2 no. 4, December 1998, pp217–228.

Franklin, S *Artificial Minds* Cambridge: MIT Press, 1995.

Minsky M *Society of Mind,* Simon and Schuster: New York, 1985

Newell A and Simon H A *Human Problem Solving,* Englewood Cliffs, NJ: Prentice-Hall, 1972.

Newell, A *Unified Theories of Cognition,* Cambridge: Harvard University Press, 1990.

Van Harmelen F and Balder J "(ML)^2: a formal language for KADS models of expertise" *Knowledge Acquisition,* 4, 121–161, 1992.

Yule P, Cooper Rand Fox J "Normative and information processing accounts of medical diagnosis" in *Proc. 20th Annual Conference of the Cognitive Science Society,* 1998.

EMOTION, INTENTION AND THE CONTROL ARCHITECTURE OF ADAPTIVELY COMPETENT INFORMATION PROCESSING

Carl B. Frankel[1] Rebecca D. Ray[2]
[1] Organizational Measurement & Engineering, San Francisco, CA, USA
[2] San Francisco State University, San Francisco, CA, USA
[1] carlf@ome1.com [2] rray@sfsu.edu

ABSTRACT

Emotionally governed, expectancy biased adaptive control is a suitable, non-conscious control architecture for the intentional processes of mind. The argument is as follows: (a) A control architecture for competent processing, expectancy biased adaptive control, is exposed. This architecture is a credible result of natural selection, and exhibits weak and strong intention. (b) The empirical literature on emotion is reviewed in terms of the expectancy biased adaptive control architecture, to argue that emotions are control signals that appraise circumstances' urgency, category, harm, benefit and uncertainty, in order to interrupt activities, regulate goal selection and modulate rate of settling. (c) The bridging concept of motivation is introduced to argue that, as control signals with the causal force to govern orderly processing in response to change, emotions supply the motive force to effect the content of intentions. (d) Reflectively conscious volition, one source of intentions but also a slow, encumbering and thus not the primary source of intentions, is one of many competing sources of demands impinging upon and resolved by the emotionally governed control system.

1 Ontology of Intention

Suppose a robot is built that knows how to secure and to use materials to replicate itself. Suppose further that one of the necessary materials is gold. The robot locates a goldmine and starts removing and smelting gold. In response, the mine's owners place barriers and hazards that the robots—many now—learn to overcome. So far, the robots are behaving like sophisticated ants. Suppose, however, that the robots start to attend to the *people* placing the obstacles. Being adaptable robots with highly resolved sensors and fast processors, they start to correlate and parse the actions that signal peoples' future actions—and peoples' deceptive falsification of signals (Ekman, 1991)—in real-time as peoples' intentions are forming, well before people know their own intentions. The robots thereby pre-empt peoples' hostile actions. Further, the same human abilities and patterns of approach and avoidance that correlate with the mine owners placing obstacles are recognized by the robots to make the owners adept and controllable miners—after all, dogs herd sheep and even some ants herd and husband aphids for the excreted sugars. The mine owners become enslaved.

These robots are like sophisticated dogs, having neither reflective consciousness nor the capacity for natural language. The robots adapt to humans' behavior, including signaling productions, in a stimulus-response, Chinese room way—that is, forming what Searle (1997) calls regulative ascription of correlation and causality rather than ascription or constitutive assignment of function. As drawn, the robots are consistent with Searle's (1992) assertion—and psychological data—that reflective consciousness is distinct from motor activity.

Yet these robots do not conform to Searle's ontological binding in which the contents of intention are inherently the product of reflective consciousness. Without reflective consciousness, the robots are displaying weak, strong and intrinsic intention (Dennett, 1996). The robots exhibit the syntax of purposiveness (weak intention), in that the content of the robots behavior is to persist in pursuing a constant outcome across varied situations using progressively efficient means. Indeed, without forming a theory of the miners' minds, the robots exhibit autonomous real-time recognition of and co-adaptation with the co-adapting intentions of others. The robots are also exhibiting the semantics of aboutness (strong intention), in that their local, gold-acquiring activity is in the service of a global goal, the content of which is to self-replicate. Finally, even if the intent to self-replicate was initially extrinsic, deriving from the robots' creators, the robots' intent is now intrinsic, since the content of creators' intentions does not include that the robots enslave people like miners to extract gold—or enslave robot makers to enhance the robots' capabilities. Each robot has an intentional, motivated mind in relation to the minds around it, as surely as a pet dog has an intentional mind in pursuit and defense of a tummy-scratch, a bone, a mate or its puppies. These robots, like pet dogs, are intentional but not reflectively conscious.

This paper presents an architecture for the intentional underpinnings of mind. The intentional contents of mind are not inherently a product of reflective, volitional consciousness but rather are any contents that become embodied in and effected by *emotional* control signals. Emotional control, and the intentional contents that emotions effect, are predominantly automatic and only sometimes influenced by conscious volition.

The argument is made in four broad strokes. (*a*) A control architecture for competent processing, expectancy biased adaptive control, is exposed. This architecture is a credible result of natural selection, and exhibits weak and strong intention. (*b*) The empirical literature on emotion is reviewed in terms of the expectancy biased adaptive control architecture, to show that emotions can credibly be conceptualized to be those control signals that appraise changing circumstances and regulate response. (*c*) The bridging concept of motivation is introduced, in order to argue that intrinsic intention and the intrinsic component of all motivation are subsumed in a single ontological category. As control signals with the causal force to govern orderly processing in response to change, *emotions supply the motive force to effect the content of intentions.* (*d*) Reflective consciousness is not needed in an architecture in which emotions motivate the realization of intentions. To the contrary, reflective processes can be slow enough as to maladaptively undermine responsiveness, were the deployment of consciousness not at the service of the control architecture. Conscious volition is one of many competing sources of demands to be resolved by the emotionally governed control system.

Emotionally governed, expectancy biased adaptive control is thus a suitable, non-conscious control architecture for the intentional processes of mind.

2 Adaptation to Stochastic Change

Independent of any role for emotion, adaptive (feedforward) control is a self-regulatory architecture that is credible to be favored by natural selection, because adaptive control competently regulates the pressures of stochastically varying circumstances in order to achieve a global goal adequately. In so doing, adaptive control exhibits strong, though not necessarily intrinsic intention. Adaptive control exhibits the co-adaptive syntax of weak intention since, with respect to immediate (local) goals, an adaptive controller can exhibit persistence of goal achievement by efficient means in varying circumstances. Adaptive control also exhibits the goal-directed semantic aboutness of strong intention, since the immediate goals of behavior occur in the service of—and thus are about—a global goal to avoid harm and to attain benefit.

2.1 The Stochastics of Competence

As basic terms, 'adaptive competence', 'information processing', 'control signal', and 'self-regulation' need definition. All four are defined in terms of a common construct, 'stochastic variation'.

One predicate for adaptive competence is a context of sufficient regularity to which to adapt. The least restrictive assumption of regularity is that the context exhibits *stochastic variation,* that is, random variation bounded by a probability distribution. If unbounded random variation is permitted, then regularity, form, order, discernible meaning and adaptation are not possible.

Adaptive competence occurs when ordered, regular and meaningfully patterned relations—standards of competence—are maintained in relation to some stochastically varying context, despite the stochastic pressure of irregularity and disorder that increases error and degrades order and discernible meaning. Adaptively competent management of stochastic variation is the problem to be solved.

Adaptive competence is an *information processing* problem, since information theory characterizes order (redundancy), and how order is preserved during processing and transmission, despite systematic, stochastic and random variation (entropy) that would degrade order. Information processing is a functional description of what an adaptively competent entity must do to manage stochastic variation.

Control signals are those which cause processing to occur at the time that it occurs. In any information processing architecture, competent control manages stochastic variation by causing correct processing to occur at a correct time or in a correct sequence.

Self-regulation names a class of information processing architectures that accomplish goals and standards (*e.g.,* standards of competence) by iterative approximation, that is, by iterative reduction of error, typically in operating environments that exhibit continuous stochastic variation. A self-regulatory architecture is therefore a natural candidate for adaptively competent information processing.

2.2 Criteria for Adaptively Competent Information Processing

Given that the problem of adaptive competence is to avoid error in the face of stochastically varying circumstances and demands, several criteria are posited that an architecture must satisfy to solve the problem of competence. The greater the stochastic pressure, *i.e.,* the lower the determinism and the greater the typical rate and direction of change, the more stringent are the criteria.

2.2.1 Adequate Correctness

Even modest stochastic pressure levies two fundament requirements for basic correctness. *Efficacy:* To avoid behavior that is random with respect to goals, each individual must form and utilize appraisals of the harms and benefits with respect to goals, and her or his efficacy with respect to avoiding harms and attaining benefits. *Synchronization:* To avoid untimely behavior, each individual must maintain adequate synchronization with circumstances, recognizing changing contingencies, and delivering responses at circumstances' rate of change.

2.2.2 Adequate Efficiency of Resource Utilization

As stochastic pressure increases, an individual not only must exhibit basic efficacy and synchronization, but also must husband her or his resources. *Economy:* To avoid endangering wastefulness, each individual must utilize resources with economy and not try impracticably both to acquire unattainable benefit and to avoid inevitable harm. *Efficiency:* To avoid problematically incomplete processing, each individual must efficiently deploy her or his limited information processing bandwidth, (*a*) by prioritizing allocation of bandwidth to most urgent events first, (*b*) by categorizing events early, to narrow the scope both of memory access and of subsequent processing, and (*c*) by consuming bandwidth for error checking only as confidence decreases and uncertainty increases. *Stability:* To avoid untenable positions, each individual must tailor her or his patterns of response to her or his adaptive niche, finding a stable position that is actuarially tenable across likely futures, and avoiding the instability of response that may lead to lethal untenability.

2.3 Expectancy-Biased Adaptive Control

Adaptive (feedforward) control (Isermann, Lachmann & Matko, 1992) is an architecture that (*a*) satisfies the stated criteria for competent response to change, even under significant stochastic pressure and (*b*) exhibits strong intention. The proposed adaptive control architecture has a local process exhibiting persistent, flexible goal attainment (syntax of weak intention). The goals of the local process are controlled by a global process that selects goals to avoid harm and attain benefit. This makes the behavior of the local process to be about achieving its goals (semantics of strong intention), since the goal-directedness is in the service of competently managing harm and benefit. By adding an expectancy bias that is experience-based, goal directedness is, in addition, about producing patterns of behavior that are competently tailored to experience of an adaptive niche.

2.3.1 Structure and Control Signaling

Adaptive control contains two, linked, self-regulatory processes (see Figure 1) and multiple control signals. *Local* to some interval in time, one self-regulatory process uses *local negative feedback* to keep current activities on track, by using goals and timetables as *local references* to assess and maintain the competence of activities' execution. *Global* to some large segment of the individual's life span, the other self-regulatory process uses *global negative feedback,* appraising realized and expected harm and benefit with respect to the *global reference* to avoid harm and to attain benefit, in order to select activities expected to prove favorable for the future.

Disparities (plus or minus) from timetable in the local process become *positive feedback events* that demand reconsideration of activities by the global process. Reconsideration by the global process can select, new goals and timetables to be *fed-forward* to the local process, as well as *dampening* of the local process' sensitivity to disparity (error).

2.3.2 Processing Narrative

The processing in adaptive control can be conceptualized and narrated as starting from a current set of goals, experiencing an interrupting contingency, selecting a response, and settling on that response.

Current goals: Local goals and standards are a (fed-forward) reference from the global process to the local process' goal-accomplishing negative feedback loop. Synchronizing timetables provide a (fed-forward) reference to the local process' positive feedback loop.

Recognition of contingency: As accomplishment of goals goes off synchronization (which includes going off goal), the positive feedback loop amplifies the disparity from timetable (with exponential gain in emergencies).

U*rgency:* Positive feedback events from the local process are an interrupting control signal to the global process. The urgency of events is signaled in both the intensity and the rate of onset of the positive feedback signal. Allocation of processing bandwidth is dynamically prioritized based on the urgency of all new events relative to that of all current activities. Lower priority events are queued, and eventually decay.

Category: For an event of sufficient priority *vis à vis* current activities, basic categorization of the event partitions memory access. Partitioning speeds processing, by reducing the amount of memory to be scanned, and by

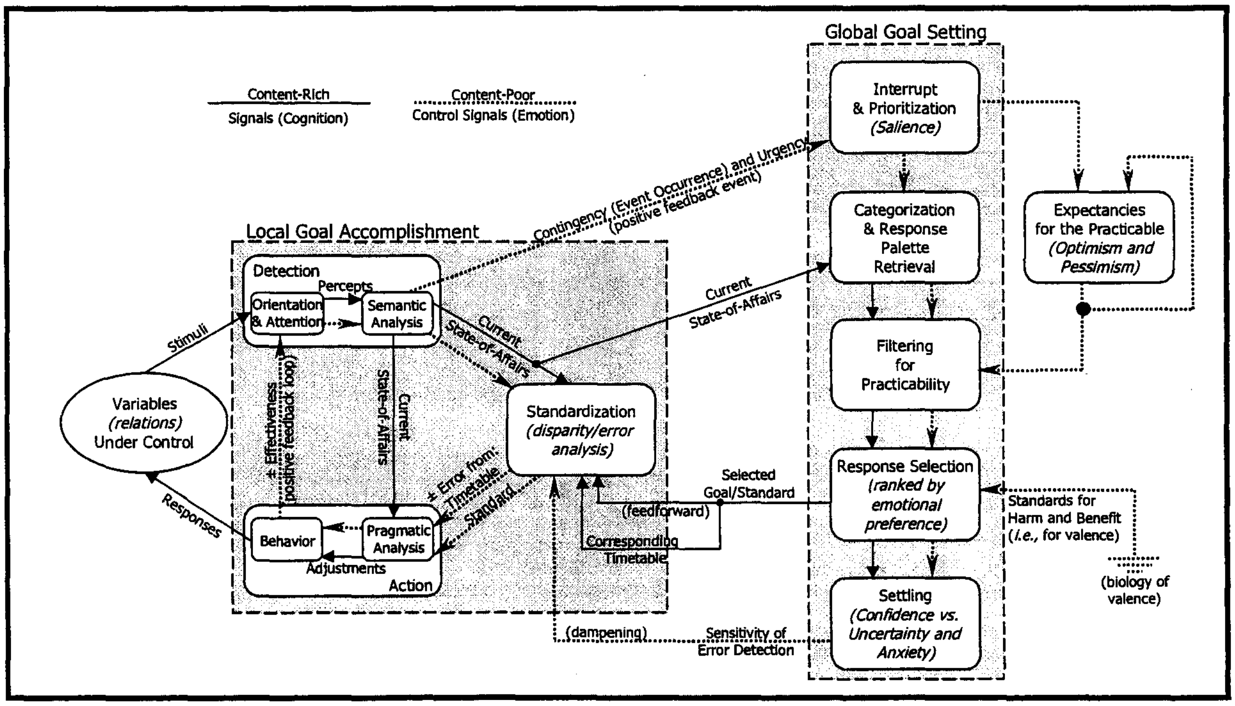

Content-Rich
Signals (Cognition)
Content-Poor
Control Signals (Emotion)
Global Goal Setting
Interrupt & Prioritization (Salience)
Categorization & Response Palette Retrieval
Expectancies for the Practicable (Optimism and Pessimism)
Filtering for Practicability
Response Selection (ranked by emotional preference)
Settling (Confidence vs. Uncertainty and Anxiety)
Standards for Harm and Benefit (i.e., for valence)
(biology of valence)
Contingency (Event Occurrence) and Urgency (positive feedback event)
Local Goal Accomplishment
Detection
Percepts
Orientation & Attention
Semantic Analysis
Current State-of-Affairs
Current State-of-Affairs
Current State-of-Affairs
Standardization (disparity/error analysis)
Stimuli
Variables (relations) Under Control
Responses
± Effectiveness (positive feedback loop)
± Error from: Timetable, Standard
Behavior
Pragmatic Analysis
Adjustments
Action
Selected Goal/Standard
(feedforward)
Corresponding Timetable
(dampening)
Sensitivity of Error Detection

Figure 1 Emotionally Governed, Expectancy Biased Adaptive Control

restricting the pallette of options retrieved for ranking.

Response evaluation: The valence of the positive feedback, positive and negative, signals current harm and benefit, which is global negative feedback with respect to the global reference to avoid harm, worst harm first, and to attain benefit. Current harm and benefit are considered in light of the bias of expected harm and benefit, in order to rank and select among response options.

Next goals: The selected response and the nominal timetable on which it should occur both feedforward reference information to the local process, giving the local process goals/standards/timetables to accomplish next.

Settling: The selected response's goodness-of-fit to circumstances, the individual's confidence in her or his ability to execute the response, and the cost of errors are all factored into the generation of a dampening signal. The dampening signal controls the local process' sensitivity to error. Sensitivity to error modulates the rate of settling on responses by controlling the amount of error-checking done, and thus also the hysteresis that controls the likelihood of further positive feedback events.

2.3.3 Expectancy Bias

The proposed adaptive control architecture is biased by expectancies for the practicable—maximum attainable benefit and minimum unavoidable harm. This pair of expectancies imposes an actuarially sound economy on patterns of response, by filtering out impracticable responses during real-time response selection: Resources are not wasted, vainly attempting to pursue what is expected to be unattainable or to avoid what is expected to be inevitable. Response is thus biased toward what is expected to be practicable and economical.

Because of their response characteristics—stability in the face of transients, some sensitivity to enduring change and zeroes that effect a reset—infinite impulse response filters (IIRs) (Hamming, 1989) model expectancies for the practicable, inputting experienced harm and benefit events. The economy consequently imposed, albeit *very* idiosyncratically tailored to the individual's adaptive niche, is typically tenable in stable niches, and stable enough to ignore transient changes. Yet IIR-expectancies are flexible enough to be responsive to some changing trends in circumstances.

3 Emotionally Governed, Expectancy Biased Adaptive Control

Biological and behavioral data suggest that *emotions are control signals* that govern response to changing circumstances, within expectancy biased adaptive control (Frankel, 1999). For an architecture of the complexity of expectancy biased adaptive control, the available empirical data on emotion are not sufficient to make a definitive case for this or any architecture. However, in addition to exhibiting strong intention and to being a kind of adequate solution to the problem of adaptive competence that natural selection would credibly favor, the proposed architecture is consistent with an abundance of data on emotion (only partially cited here). The proposed architecture thus makes a plausible, *constructive* case for emotion as a control signal in the production of competent, intentional behavior.

3.1 Defining Emotion

In psychology there is no agreement on how to define 'emotion', nor is there a single superordinate term that covers all of the phenomena that are in some sense emotional (Gross, 1998). Adding complexity, most psychologists use a naive taxonomy of psychological phenomena in which 'cognition' and 'information processing' are interchangeable, implicitly relegating emotion to a role where it can neither contain information nor be a descriptive or inferential process.

Herein, 'emotion' is used as the superordinate term, applicable to all realized and expected valence states that typically appraise harm and benefit. Realized valence that appraises realized harm and benefit is realized emotion. Expected valence that appraises expected harm and benefit is expected emotion. This is not to say that emotion encodes only valence; to the contrary, emotions typically encode many forms of appraisal, as described below. Rather, valenced appraisal of harm and benefit is the defining characteristic of (necessary and sufficient for) emotion. Emotions include valence states of any duration, from micro-momentary to lifespan. Emotions also include valence states of any abstraction, from hunger and pain, through fear, anger and happiness, to embarrassment, malaise and *ennui*.

3.2 Emotions As Control Signals

There are not yet definitive neurological data to show that emotions control orderly processing of change, since that would require a (non-existent) complete map of the brain showing the necessity of emotion throughout neural control of change processing. However, the available neurological data strongly support the necessity of emotion to be a reasonable hypothesis.

In two regions of the brain that are very separate, both physically and functionally, dramatic degradation of motivation and organization is observed when the brain's capacity for orderly emotional processing is damaged. (*a*) Lesions that include the amygdala and some surrounding tissue can flatten emotional response and disable the regulation of attention, disrupting the process of salience (LeDoux, 1992). (*b*) Lesions to the prefrontal lobes that disable the operation of emotions also disable the organization of motivation and behavior (Damasio, 1994). People with prefrontal lesions behave exactly like programmed control systems with control signal failure: They have procedural knowledge intact, but without operational emotions, cannot provide real-time control to behavior to execute knowledge.

The fact of two independent points of failure militates against coincidental co-location of emotion and control. While not in itself definitive evidence for the necessity of emotion for control, it is strong evidence for the reasonableness of a hypothesis of necessity.

3.3 Emotional Governance

In real-time, emotions appraise change and govern response, acting as control signals that (*a*) interrupt current activity and prioritize processing, (*b*) categorize events, (*c*) filter response options for expected practicability, (*d*) rank options for a favorable future and (*e*) modulate settling on the ranking response.

3.3.1 Emotional Onset: Urgency of Contingencies and Consequent Prioritized Interrupts

In order to interrupt current activities and to allocate processing bandwidth, emotions signal the relative urgency and priority of events in both emotional intensity and rate of emotional onset (on Figure 1, the positive feedback loop, the Contingencies control flow and the Salience process). Going off-timetable (including off-goal) is a contingency that triggers emotions (Carver & Scheier, 1990). Emotions raise alertness, altering the breadth of attention as needed (Derryberry & Tucker, 1994). Emotions orient attention to focus on change (Posner & Petersen, 1990). Finally, emotions begin to assess the change as potentially harmful or beneficial, and arouse the individual both autonomically and behaviorally (Lang, Bradley & Cuthbert, 1990). As a result, objects of attention take on sustained salience as a contingency interrupting activities in a prioritized way.

Dynamic Prioritization: Events of unequal priority yield a single emotion that reflects and focuses on the event of highest priority (Frijda, 1988). All lower priority events are queued, to be serviced after higher priority events, or else decaying from the queue as emotions decay. Events of equal priority can result in multiple, simultaneous "mixed" and possibly conflicting emotions.

3.3.2 Emotional Categorization: Memory Partitioning

After initial emotional onset, interrupt and shift of attention, emotions categorize events, signaling what kind of an event each event is, thereby partitioning memory, so as to access and process an appropriate palette of response options (on Figure 1, the Categorization process). While not yet converging on specific, neurologically embodied categories, many investigators agree that emotions categorize events (Ekman, 1992; Gray, 1990; Panksepp, 1989). Emotion categories are understood to reflect action tendencies (Frijda, Kuipers & ter Schure, 1989) rather than ballistic trajectories, because emotions decouple the contents of attention from otherwise reflexive response (Scherer, 1994). Emotional categories thus do not lead to rigidly stereotyped behaviors. Rather, once attention is focused, emotional categories prime memory so that a palette of responses is quickly retrieved, based upon responses' coherence with prevailing emotional appraisals of current circumstances.

3.3.3 Emotional Expectancy Biasing: Insuring Practicability

After categorization, the resultant palette of response options is biased to eliminate wastefully impracticable options that try either to attain what is expected to be unattainable positive emotion or to avoid what is expected to be inevitable negative emotion—and by proxy, unattainable benefit and inevitable harm (on Figure 1, the Expectancy and Practicability processes). The biasing filters are expectancies for maximum attainable positive emotion and minimum unavoidable negative emotion—proxies for maximum attainable benefit and minimum unavoidable harm (see section 4.2).

3.3.4 Emotional Valuation: Selecting Efficacious Responses

During response selection, emotional valence and intensity, expected and realized, control the ranking of the biased palette of options. (on Figure 1, Response Selection process, and Goal and Timetable feedforward control flows). The standard for ranking is emotionally risk-averse: *To avoid what are expected to be negative emotions, worst emotions first, and then to pursue what are expected to be positive emotions.* The ranking option and its timetable are selected and fed-forward.

Utility: Neurologically, the evaluation of stimuli (Davidson, 1992) and utility (Ito & Cacioppo, 1999) is encoded in emotional valence, biased toward aversion to the risk of negative emotions (Ito, Larsen, Smith & Cacioppo, 1998). Behaviorally, in ranking the utility of harm and benefit, contrary to prospect theory (Kahneman & Tversky, 1990), people avoid the negative emotion of regret associated with a loss, not the loss *per se* (Larrick, 1993). Regret avoidant options may be either risk avoidant or risk taking (Zeelenberg & van Dijk, 1997).

Negative emotion aversion: Regret is not the only strong aversion. Before people accept helplessness, they exhibit reactance (Brehm & Sensenig, 1966). Avoidance of anxiety is a powerful motivator (Greenberg, Pyszczynski, & Solomon, 1995). Shame avoidance increases aggression and narrows peoples' focus so that they do not take the perspectives of others, harming relationships (Tangney, Wagner, Hill-Barlow, Marschall & Gramzow, 1996). Abandonment and betrayal are also worst emotions that people typically avoid systematically.

Automaticity: The avoidance of worst emotions is often so automatic and so successful as often to occur completely outside of consciousness. For example, when people get dressed to go out at the start of their day, most do not give any conscious attention, thought or feeling to the fact that they are doing so, in part, to avoid the shame of going naked in a clothed world. Yet most people are immediately alarmed and avoidant at the suggestion.

3.3.5 Emotional Dampening: Confidence & Settling

Automaticity is the special case of settling, where the individual is fully confident that a selected response is beyond the possibility of error. More generally, once a response option has been selected and fed forward, the rate of response settling is modulated by the individual's emotional confidence *vs.* anxiety (on Figure 1, the Settling process and Sensitivity Dampening control flow).

Confidence: The individual's level of confidence reflects her or his belief (a) that the selected response is a certain fit to circumstances, (b) that the task difficulty is within capabilities and (c) that the cost of likely errors is affordable. The greater the individual's confidence, the more certain and compelling is the response, and thus the more dampened the individual's sensitivity to error and the more efficient the settling. After people select how to respond, their natural predilection is to be confident that they can implement their decision successfully (Taylor & Gollwitzer, 1995).

Uncertainty and anxiety: As the selected response is a poor or uncertain fit, as the response taxes abilities, or as the cost of errors increases, confidence lowers and anxiety increases. The greater the individual's anxiety and uncertainty, the less dampened is the sensitivity to error and either the slower and less efficient the settling, or else the more erratic the settling as time pressure increases. Anxiety reflects uncertainty (Epstein & Roupenian, 1970; Feather, 1963, 1965; Wright, 1984). Tolerable levels of both uncertainty (Siegman & Pope, 1965) and anxiety (Gray, 1990) slow settling. As stress increases, people make and consider fewer distinctions, rushing to settle before they have considered all available alternatives (Keinan, Friedland & Arad, 1991).

Settling strategies: Life is often very uncertain, and errors often costly, militating against easy settling. Yet competent settling demands a dampening function that modulates settling to match circumstances' rate of change. Failure to settle in time is often catastrophic, making it credible that natural selection would favor a design that creates punishing internal pressure to settle.

Faced with a punishing emotional dampening mechanism, people compromise on a preferred settling style. Sorrentino (*e.g.*, Sorrentino, Holmes, Hanna & Sharp, 1995) has found that some people ignore anxiety-raising discrepancies, settling rapidly, even prematurely, with certainty and confidence, and cleaving their social universe into trustworthy or not. Others have evenly modulated anxiety, error-checking and settling, taking in more information and subjecting it to more careful scrutiny, but seldom establishing a more than moderately trusting position.

Still others stay chronically anxious and inefficient, settling erratically. The chronically anxious prefer a narrow focus (Stoeber, 1996) on possible error at the expense of sometimes-important information. Anxious focus is biased toward the processing of threat, much of which is minor in nature, to which anxious people are more attentive, by which they are more distracted (McNally, 1996) and about which they ruminate. Worriers have low tolerance for uncertainty, are disproportionately sensitive to uncertainty, and expect uncertainty to bring failure (Shimkunas, 1970).

3.3.6 Coping and Emotion Repair

As demands increase, people increase their problem solving output to keep pace. Eventually, however, people reach a point where they are consistently too wrong or too late or both. People reach a breaking point, a positive feedback event where they recognize that they cannot keep pace, or where they decide that the costs exceed the benefits. At that breaking (inflection) point, the adaptive strategy shifts from problem solving to coping. People disengage from the focal problem, and either start addressing peripheral problems that staunch their loss of ground, or start directly repairing their negative emotion, or both. Moreover, if being pushed past the breaking point is accompanied by a concomitant shift of expectancies, people may not notice after demands decrease, and may resist re-engagement with problems.

3.3.7 Metastable Equilibrium - Failure to Settle

Adaptive control designs are vulnerable to metastable equilibrium. For people, goal conflict can produce this kind of failure to settle. The immediate result of conflicting demands and mixed emotions is increased stress, slowed response and high error rate (Smith & Gehl, 1974). Mixed emotional states are stressful and disruptive, and when sustained, result in high levels of negative emotion and psychosomatic complaints. Such ambivalent states demand substantial bandwidth to process, and stymie action (Emmons & King, 1988). Conflicting standards result in increased distractibility, uncertainty, and indecisiveness, thereby disorganizing motivation (van Hook & Higgins, 1988). People can panic in the face of irreducible goal conflict, producing a rush to settle; however, anxiety may also inhibit panic, creating paralysis (Gray & McNaughton, 1996). Unresolvable or irreconcilable demands are both seriously disorganizing and highly dysphoric.

3.4 Bias by Emotional Expectancies

To be actuarially tenable, patterns of response should be tailored to be (*a*) adequate to the largest range of likely futures in a given adaptive niche, (*b*) insensitive to transient changes in the niche, and (*c*) sensitive to changing trends in the niche. Responses are biased toward tenability by IIR filters (Optimism and Pessimism in Figure 1) that sample emotional valence events. IIRs formulate expectancies that ignore most transients and tracks some trends. As a result, the individual's emotional expectancies for the bounds of the emotionally practicable comprise a stable, idiosyncratic biasing to the individual's adaptive niche, reflecting her or his unique emotional experience, education and acculturation.

Valence expectancy is usually a cognitive construct, *e.g.*, self-esteem, possible self, ideal *vs.* ought self, prevention *vs.* promotion focus, or dispositional optimism *vs.* pessimism. However, all of these valenced constructs are predicated upon a common pair of underlying emotional expectancies. Maximum attainable benefit is the expectancy for the threshold beyond which benefit and positive emotion are not practicably attainable. Minimum unavoidable harm is the expectancy for the threshold below which harm and negative emotion are inevitable, *vs.* worse, avoidable harm and emotion.

People maintain expectancies for both positive and negative emotion (Marshall, Wortman, Kusulas, Hervig. & Vickers, 1992), each with a distinct neurological basis

(Davidson, 1993) Emotional expectancy comprises an assessed emotional trend, predicted from emotional events whenever they occur in an interval, with discrepant samples being ignored if they do not reflect the kind of trend that signals possible enduring change (Varey & Kahneman, 1992). Consistent with the smoothing of IIR output, emotional output is stably positive and negative over long intervals of time (Watson & Clark, 1984).

3.4.1 Stable Patterns of Emotional Response

Emotional expectancies stabilize patterns of appraisal. Emotional expectancies smooth emotions toward expected values (Wilson, Lisle, Kraft, & Wetzel, 1989), direct attention toward expectancy-consistent stimuli (Byrne & Eysenck, 1995), accept expectancy-consistent emotions as informative and reject expectancy-inconsistent emotions as noise (Gaspar & Clore, 1998), and disambiguate ambiguities and assess performance outcomes toward expectations (Brown & Dutton, 1997).

Emotional expectancies also stabilize patterns of emotional and behavioral response. People with high negative affectivity tend to experience stable discomfort, independent of time, situation or identifiable stressors (Watson & Walker, 1996). Pessimists tend to expect to feel worse, to experience lower life satisfaction and more depressive symptoms (Chang, Maydeu-Olivares & D'Zurilla, 1997) and to be more vulnerable to making negative self-assessments (Brown & Mankowski, 1993). The converse is true for optimistic people.

Emotional expectancies are often self-reinforcing. Optimists tend to stay socially engaged and focused on hopeful aspects of circumstances, while pessimists are likely to focus on stressful aspects of circumstances and to disengage from problems (Scheier, Weintraub & Carver, 1986). Keeping resources focused on problems for longer, an optimistic strategy is stochastically more likely to produce solutions and expectancy-reinforcing positive emotion. The pessimist withdraws resources sooner, increasing the risk of failure and expectancy-reinforcing negative emotion.

Emotional expectancies can be so stable and self-reinforcing that idiosyncratic patterns of response, tailored to one adaptive niche, often persist when the niche changes or when the individual is transplanted to another niche. Miscontextualized adaptations and coping strategies often persevere as overly stable, even rigidly psychopathological, individual differences. Although individual competence is not best served by such rigidity, the species' genetic fitness can benefit. The broad pallette of individuals' strategies available at any point in time increases the likelihood that some individuals will be well suited to new circumstances, when circumstances change.

3.4.2 Flexibility in Response to Changing Trends

Emotional expectancies for the practicable can change consistent with an IIR construction. IIRs can respond selectively to enduring change. IIRs also have regions of reset (zeroes), where surprise can make emotional expectancies change abruptly.

When dramatic life change results in enduringly different emotions, patterns of emotional expectancy can change. For example, falling in love heightens positive emotional expectancy—which typically then decays as romance cools and expectancies are not refreshed with enduring, strong positive emotions. Traumatic events and their sequellae often generate enduring emotional change that heightens negative emotional expectancies.

Surprise accompanied by sustained interest resets expectancies (that is, surprise is a zero of the IIR, driving IIR output to zero, no expectancy). Thereafter, expectancies assume values from post-surprise emotional events. At onset, the surprising stimulus is persistently salient (Meyer, Niepel, Rudolph, & Schuetzwohl, 1991). Processing slows, as people allocate processing resources for an attributional search (Stiensmeier-Pelster, Martini & Reisenzein, 1995). If attribution fails, one of three outcomes occurs. (a) The uninterpretable event is deemed unimportant and is ignored. (b) The uninterpretable event is deemed to have potentially catastrophic significance, provokes significant anxiety, and a defense is quickly settled upon. (c) An event that is deemed important but not catastrophically threatening, provokes at most tolerable anxiety and also sustained interest. This third type of surprise event, a "disturb-then-reframe" protocol, causes expectancies to take on new values (Davis & Knowles, 1999). Surprise and interest may also promote change in psychotherapy (Omer, 1990). The growing trust in a therapeutic alliance can be understood both to increase sensitivity to emotion by lowering the noise of anxiety, and to increase the tolerability of emotion, thus stochastically increasing the likelihood of transformative surprise events in treatment.

4 Ontological Binding of Intention to Motivation and Emotion

The proposed adaptive control architecture exhibits strong intention, but not necessarily intrinsic intention. Emotional control signaling, by contrast, is a fundamental intrinsic of human information processing. To complete the argument that emotions automate and effect the content of intentions, emotions and intentions must be linked. The bridging concept between intention and emotion is motivation. The causal force by means of which emotions govern behavior and effect intrinsic intention is *motive force.*

While not agreeing on the determinants of motivation, psychologists generally agree on the necessity of motivation: Without motivation, competently organized behavior is unlikely to occur on a sustained basis. While much motivation has extrinsic determinants, this paper takes the position that all motivation has a necessary intrinsic component that appraises the significance of extrinsic factors, in order to control the organization of behavior consistent with the content of intrinsic signification. For example, confronted with an extrinsic like a snake during a stroll, most people will be motivated to step around it, whereas a phobic might be motivated to leave the area, while a herpetologist might be motivated to pick up the snake and study it. To be realized, all motivation is implemented by an intrinsic motive force that effects the contents of intrinsic signification.

This decomposition of motivation suggests that the intrinsic component of motivation and intrinsic intention comprise a single ontological category. *Organization of behavior:* Both weak intention and intrinsic motivational components result in organized behavior with respect to changing circumstances. *Content of motivation/intention:* The content of a strong intention's aboutness is a motive's intrinsic significance. To say that a person is motivated or intends to step around a snake is to say that the detour is about avoiding the snake. *Intrinsic locus:* Both intrinsic intention and the intrinsic component of motivation assure aboutness rather than tropism. Stepping around the snake is both motivated and intentional behavior about avoiding the snake, because avoiding the snake is in the service of an intrinsic motive and intention, *viz.*, avoiding harms and attaining benefits as the individual construes them.

Emotions are ontologically bound to intrinsic intention/motivation, because emotions are the control signals that appraise intrinsic significance and effect the contents of intrinsic intention/motivation. *Content of emotion:* From the onset of change to settling on a response, emotions appraise the significance of changing circumstances, in terms of their urgency, category, harm benefit, and uncertainty. From these appraisals, emotions organize cognition and behavior to be about avoidance of harms and attainment of benefits. *Intrinsic locus:* Emotions organize cognition and behavior in the service of an intrinsic motive/intention: To avoid what are expected to be negative emotions, worst emotions first, and to attain what are expected to be positive emotions. To the extent that emotions, realized and expected, accurately appraise harms and benefits, realized and expected, the favorable regulation of future emotions regulates future competence by proxy. Emotions, as internal control signals with causal force, automate and effect the contents of intentions with motive force.

5 Primacy of Emotion for Intention

To perform competently under typical stochastic pressures, it is critical that control systems have the ability first to be on time and then secondarily to be as accurate as possible, since being too late is often as catastrophic a form of being too wrong as is being too inaccurate. Slower forms of processing that depend on high level data abstractions, *e.g.*, symbols, propositions, metaphors, modals, must be separable from, be secondary to and operate at the service of control. Therefore, in the proposed control architecture, cognitive components operate under emotional control, at the service of the intentions that emotions effect and automate. Language, planning and reflective consciousness serve the global, emotionally controlled goal to avoid what are expected to be negative emotions, worst emotions first, and to pursue what are expected to be positive emotions.

With its flexible context (option) generation, its insight into the distant future, and as keeper of the broader social and moral contract, conscious volition can sometimes overcome an immediate and short-sighted impulse by injecting internal percepts of long term consequences, both of the impulse and of alternative behavioral pathways. However, if, relative to other contingencies impinging on the emotional control system, volitional percepts do not invoke emotions of sufficient priority to hold attention and to motivate behavior, consciousness is of little controlling effect. Reflective consciousness and its volition are secondary, modulator functions. The intrinsic intention to use current and expected emotions favorably to regulate future emotions, and by proxy future competence, is primary.

6 Conclusion

Confronted with stochastically varying circumstances, the human mind is, of adaptive necessity, primarily a control system. Substantial recent data suggest that emotionally governed, expectancy biased adaptive control is a suitable control architecture. The human mind is thereby competent as a result of being motivated by experienced and expected emotions favorably to regulate future emotions, and by proxy adaptive competence. Emotionally governed goal accomplishment exhibits the co-adaptive syntax of weak intention. Emotional governance also exhibits the aboutness semantics of intrinsic strong intention, because goals serve the intrinsic standard to avoid negative emotions and to attain positive emotions.

Largely automatically, emotions govern the human mind's information processing with motive force, controlling salience, priority, patterns of response, confidence and disposition so as to co-adapt with changing circumstances. Favored by natural selection—both because (*a*) emotions typically position individuals adequately competently and because (*b*) emotions' idiosyncrasy promotes individual differences, creating a broad, risk-reducing pool of strategies for the species—emotions are control signals that govern the regulation of behavior and future emotions. Emotions mediate, motivate and organize adaptive competence, such that individuals avoid harms and attain benefits as their emotions appraise them. Emotions thereby automate, realize and signal the contents of the mind's intentions.

References

Brehm, J. W. & Sensenig, J. (1966) Social influence as a function of attempted and implied usurpation of choice. *Journal of Personality and Social Psychology, 4:* 703-707.

Brown, J. D. & Dutton, K. A. (1997) Global self-esteem and specific self-views as determinants of people's reactions to success and failure. *Journal of Personality and Social Psychology, 73:* 139-148.

Brown, J. D. & Mankowski, T. A. (1993) Self-esteem, mood, and self-evaluation: Changes in mood and the way you see you. *Journal of Personality and Social Psychology, 64:* 421-430.

Byrne, A. & Eysenck, M. W. (1995) Trait anxiety, anxious mood, and threat detection. *Cognition and Emotion, 9:* 549-562.

Carver, C. S. & Scheier, M. F. (1990) Origins and functions of positive and negative affect: A control-process view. *Psychological Review, 97:* 19-35.

Chang, E. C., Maydeu-Olivares, A. & D'Zurilla, T. J. (1997) Optimism and pessimism as partially independent constructs: Relationship to positive and negative affectivity and psychological well-being. *Personality & Individual Differences, 23:* 433-440.

Damasio, A. R. (1994) *Descartes' error: Emotion, reason, and the human brain.* New York: G.P. Putnam.

Davidson, R. J. (1992). Prolegomenon to the structure of emotion: Gleanings from neuropsychology. *Cognition and Emotion, 6:* 245-268.

Davidson, R. J. (1993) Parsing affective space: Perspectives from neuropsychology and psychophysiology. *Neuropsychology, 7:* 464-475.

Davis, B. P. & Knowles, E. S. (1999) A disrupt-then-reframe technique of social influence. *Journal of Personality and Social Psychology, 72:* 192-199.

Dennett, D. C. (1996) *Kinds of minds.* New York: Basic Books.

Derryberry, D. & Tucker, D. (1994) Motivating the focus of attention. In P. M. Niedenthal & S. Kitayama (Eds.), *The heart's eye: Emotional influences in perception and attention,* (pp. 167-196). San Diego, CA: Academic Press.

Ekman, P. (1991) *Telling lies: Clues to deceit in the marketplace, politics and marriage.* New York: W.W. Norton.

Ekman, P. (1992) An argument for basic emotions. *Cognition and Emotion, 6:* 169-200.

Emmons, R. A. & King, L. A. (1988) Conflict among personal strivings: Immediate and long-term implications for psychological and physical well-being. *Journal of Personality and Social Psychology, 54:* 1040-1048.

Epstein, S. & Roupenian, A. (1970) Heart rate and skin conductance during experimentally induced anxiety: The effect of uncertainty about receiving a noxious stimulus. *Journal of Personality and Social Psychology, 16:* 20-28.

Feather, N. T. (1965) The relationship of expectation of success to need achievement and test anxiety. *Journal of Personality and Social Psychology, 1:* 118-126.

Feather, N. T. (1963) The effect of differential failure on expectation of success, reported anxiety, and response uncertainty. *Journal of Personality, 31:* 289-312.

Frankel, C. B. (1999) Such order from confusion sprung: Affect regulation and adaptive competence. Dissertation presented to the faculty of Pacific Graduate School of Psychology, Palo Alto, CA, USA

Frijda, N. (1988) The laws of emotion. *American Psychologist, 43:* 349-358.

Frijda, N., Kuipers, P. & ter Schure, E. (1989) Relations among emotion, appraisal, and emotional action readiness. *Journal of Personality and Social Psychology, 57:* 212-228.

Gasper, K. & Clore, G. L. (1998) The persistent use of negative affect by anxious individuals to estimate risk. *Journal of Personality and Social Psychology, 74:* 1350-1363.

Gray, J. A. (1990) Brain systems that mediate both emotion and cognition. *Cognition and Emotion, 4:* 269-288.

Gray, J. A. & McNaughton, N. (1996) The neuropsychology of anxiety: Reprise. In D. Hope (Ed.), *Perspective on anxiety, panic and fear,* (pp. 61-134). Omaha, NB: University of Nebraska Press.

Greenberg, J., Pyszczynski, T. & Solomon, S. (1995) Toward a dual-motive depth psychology of self and social behavior. In M. H. Kernis (Ed.), *Efficacy, agency, and self-esteem,* (pp. 73-99). New York: Plenum Press.

Gross, J. J. (1998) The emerging field of emotion regulation: An integrative review. *Review of general psychology, 2:* 271-299.

Hamming, R. W. (1989) *Digital filters.* Englewood Cliffs, NJ: Prentice Hall.

Isermann, R., Lachmann, K.-H. & Matko, D. (1992) *Adaptive control systems.* New York: Prentice Hall.

Ito, T. A. & Cacioppo, J. T. (1999) The psychophysiology of utility appraisals. In D. Kahneman, E. Diener & N. Schwarz, (Eds.), *Well-being: The foundations of hedonic psychology,* (pp. 470-488). New York: Russell Sage Foundation.

Ito, T. A., Larsen, J. T., Smith, N. K. & Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: The negativity bias in evaluative categorizations. *Journal of Personality and Social Psychology, 75:* 887-900.

Kahneman, D. & Tversky, A. (1990) Prospect theory: An analysis of decision under risk. In P. K. Moser (Ed.), *Rationality in action: Contemporary approaches,* (pp. 140-170). New York: Cambridge University Press.

Keinan, G., Friedland, N. & Arad, L. (1991) Chunking and integration: Effects of stress on the structuring of information. *Cognition and Emotion, 5:* 133-145.

Lang, P. J., Bradley, M. M. & Cuthbert, B. N. (1990) Emotion, attention, and the startle reflex. *Psychological Review, 97:* 377-395.

Larrick, R. P. (1993) Motivational factors in decision theories: The role of self-protection.*Psychological Bulletin, 113:* 440-450.

LeDoux, J. E. (1993) Emotion & the amygdala. In J. P. Aggleton (Ed.), *The amygdala: Neurobiological aspects of emotion, memory, & mental dysfunction,* (pp. 339-351). New York: Wiley-Liss.

Marshall, G. N., Wortman, C. B., Kusulas, J. W., Hervig, L. K. & Vickers, R. R., Jr. (1992) Distinguishing optimism from pessimism: Relations to fundamental dimensions of mood and personality. *Journal of Personality and Social Psychology, 62:* 1067-1074.

McNally, R. J. (1996) Cognitive bias in anxiety disorders. In D. Hope (Ed.), *Perspectives on anxiety, panic and fear,* (pp. 213-250). Omaha, NB: University of Nebraska Press.

Meyer, W. U., Niepel, M., Rudolph, U. & Schuetzwohl, A. (1991) An experimental analysis of surprise. *Cognition and Emotion, 5:* 295-311.

Omer, H. (1990) Enhancing the impact of therapeutic interventions. *American Journal of Psychotherapy, 44:* 218-231.

Panksepp, J. (1989) The neurobiology of emotions: Of animal brains and human feelings. In H. Wagner & A. Manstead (Eds.), *Handbook of social psychophysiology,* (pp. 5-26). Chichester, UK: Wiley.

Posner, M. I. & Petersen, S. E. (1990) The attention system of the human brain. *Annual Review of Neuroscience, 13:* 25-42.

Scheier, M. F., Weintraub, J. K. & Carver, C. S. (1986) Coping with stress: Divergent strategies of optimists and pessimists. *Journal of Personality and Social Psychology, 51:* 1257-1264.

Scherer, K. R. (1994) Emotion serves to decouple stimulus and response. In P. Ekman and R. J. Davidson (Eds.), *The nature of emotions: Fundamental questions,* (pp.127-130). New York: Oxford University Press.

Searle, J. R. (1992) *The rediscovery of mind.* Cambridge, MA: MIT Press.

Searle, J.R. (1997) *The Construction of social reality.* New York: Free Press.

Siegman, A. W. & Pope, B. (1965) Effects of question specificity and anxiety-producing messages on verbal fluency in the initial interview. *Journal of Personality and Social Psychology, 2:* 522-530.

Shimkunas, A. M. (1970) Anxiety and expectancy change: The effects of failure and uncertainty. *Journal of Personality and Social Psychology, 15:* 34-42.

Smith, B. D. & Gehl, L. (1974) Multiple-exposure effects of resolutions of four basic conflict types. *Journal of Experimental Psychology, 102:* 50-55.

Sorrentino, R. M., Holmes, J. G., Hanna, S. E. & Sharp, A. (1995) Uncertainty orientation and trust in close relationships: Individual differences in cognitive styles. *Journal of Personality and Social Psychology, 68:* 314-327.

Stiensmeier-Pelster, J., Martini, A & Reisenzein, R. (1995) The role of surprise in the attribution process. *Cognition and Emotion, 9:* 5-31.

Stoeber, J. (1996) Anxiety and the regulation of complex problem situations: Playing it safe? In W. Battmann & S. Dutke (Eds.), *Processes of the molar regulation of behavior,* (pp.105-118). Scottsdale, AZ: Pabst Science Publishers.

Tangney, J. P., Wagner, P. E., Hill-Barlow, D., Marschall, D. E. & Gramzow, R. (1996) Relation of shame and guilt to constructive vs. destructive responses to anger across the lifespan. *Journal of Personality and Social Psychology, 70:* 797-809.

Taylor, S. E. & Gollwitzer, P. M. (1995) Effects of mindset on positive illusions. *Journal of Personality and Social Psychology, 69:* 213-226.

van Hook, E. & Higgins, E. T. (1988) Self-related problems beyond the self-concept: Motivational consequences of discrepant self-guides. *Journal of Personality and Social Psychology, 55:* 625-633.

Varey, C. A. & Kahneman, D. (1992) Experiences extended across time: Evaluation of moments and episodes. *Journal of Behavioral Decision Making, 5:* 169-185.

Watson, D. & Clark, L. A. (1984) Negative affectivity: The disposition to experience aversive emotional states. *Psychological Bulletin, 96:* 465-490.

Watson, D. & Walker, L. M. (1996) The long-term stability and predictive validity of trait measures of affect. *Journal of Personality and Social Psychology, 70:* 567-577.

Wilson, T. D., Lisle, D. J., Kraft, D. & Wetzel, C. G. (1989) Preferences as expectation-driven inferences: Effects of affective expectations on affective experience. *Journal of Personality and Social Psychology, 56:* 519-530.

Wright, R. A. (1984) Motivation, anxiety, and the difficulty of avoidant control. *Journal of Personality and Social Psychology, 46:* 1376-1388.

Zeelenberg, M. & van Dijk, E. (1997) A reverse sunk cost effect in risky decision making: Sometimes we have too much invested to gamble. *Journal of Economic Psychology, 18,* 677-691.

A "Consciousness" Based Architecture for a Functioning Mind

Stan Franklin

Institute for Intelligent Systems and
Department of Mathematical Sciences
The University of Memphis
stan.franklin@memphis.edu

Abstract

Here we describe an architecture designed to accommodatemultiple aspects of human mental functioning. In a roughly star-shaped configuration centered on a "consciousness" module, the architecture accommodates perception, associative memory, emotions, action-selection, deliberation, language generation, behavioral and perceptual learning, self-preservation and metacognition modules. The various modules (partially) implement several different theories of these various aspects of cognition. The mechanisms used in implementing the several modules have been inspired by a number of different "new AI" techniques. One software agent embodying much of the architecture is in the debugging stage (Bogner et al. in press). A second, intending to include all of the modules of the architecture is well along in the design stage (Franklin et al. 1998). The architecture, together with the underlying mechanisms, comprises a fairly comprehensive model of cognition (Franklin & Graesser 1999). The most significant gap is the lack of such human-like senses as vision and hearing, and the lack of real-world physical motor output. The agents interact with their environments mostly through email in natural language.

The "consciousness" module is based on global workspace theory (Baars 1988, 1997). The central role of this module is due to its ability to select relevant resources with which to deal with incoming perceptions and with current internal states. Its underlying mechanism was inspired by pandemonium theory (Jackson 1987).

The perception module employs analysis of surface features for natural language understanding (Allen 1995). It partially implements perceptual symbol system theory (Barsalou 1999), while its underlying mechanism constitutes a portion of the copycat architecture (Hofstadter & Mitchell 1994).

Within this architecture the emotions play something of the role of the temperature in the copycat architecture and of the gain control in pandemonium theory. They give quick indication of how well things are going, and influence both action-selection and memory. The theory behind this module was influenced by several sources (Picard 1997, Johnson 1999, Rolls 1999). The implementation is via pandemonium theory enhanced with an activation-passing network.

The action-selection mechanism of this architecture is implemented by a major enhancement of the behavior net (Maes 1989). Behavior in this model corresponding to goal contexts in global workspace theory. The net is fed at one end by environmental and/or internal state influences, and at the other by fundamental drives. Activation passes in both directions. The behaviors compete for execution, that is, to become the dominant goal context.

The deliberation and language generation modules are implemented via pandemonium theory. The construction of scenarios and of outgoing messages are both accomplished by repeated appeal to the "consciousness" mechanism. Relevant events for the scenarios and paragraphs for the messages offer themselves in response to "conscious" broadcasts. The learning modules employ case-based reasoning (Kolodner 1993) using information gleaned from human correspondents. Metacognition is based on fuzzy classifier systems (Valenzuela-Rendon 1991).

As in the copycat architecture, almost all of the actions taken by the agents, both internal and external, are performed by codelets. These are small pieces of code typically doing one small job with little communication between them. Our architecture can be thought of as a multi-agent system overlaid with a few, more abstract mechanisms. Altogether, it offers one possible architecture for a relatively fully functioning mind. One could consider these agents as early attempts at the exploration of design space and niche space (Sloman 1998).

Autonomous Agents

Artificial intelligence pursues the twin goals of understanding human intelligence and of producing intelligent software and/or artifacts. Designing, implementing and experimenting with autonomous agents furthers both these goals in a synergistic way. An *autonomous agent* (Franklin & Graesser 1997) is a system situated in, and part of, an environment, which senses that environment, and acts on it, over time, in pursuit of its own agenda. In biological agents, this agenda arises from evolved in drives and their associated goals; in artificial agents from drives and goals built in by its creator. Such drives, which act as motive generators (Sloman 1987), must be present, whether explicitly represented, or expressed causally. The agent also acts in such a way as to possibly influence what it senses at a later time. In other words, it is structurally coupled to its environment (Maturana 1975, Maturana et al. 1980). Biological examples of autonomous agents include humans and most animals. Non-biological examples include some mobile robots, and various computational agents, including artificial life agents, software agents and many computer viruses. We'll be concerned with autonomous software agents, designed for specific tasks, and 'living' in real world computing systems such as operating systems, databases, or networks.

Global Workspace Theory

The material in this section is from Baars' two books (1988, 1997) (1988, 1997) and superficially describes his global workspace theory of consciousness.

In his global workspace theory, Baars, along with many others (e.g. (Minsky 1985, Ornstein 1986, Edelman 1987)) , postulates that human cognition is implemented by a multitude of relatively small, special purpose processes, almost always unconscious. (It's a multiagent system.) Communication between them is rare and over a narrow bandwidth. Coalitions of such processes find their way into a global workspace (and into consciousness). This limited capacity workspace serves to broadcast the message of the coalition to all the unconscious processors, in order to recruit other processors to join in handling the current novel situation, or in solving the current problem. Thus consciousness in this theory allows us to deal with novelty or problematic situations that can't be dealt with efficiently, or at all, by habituated unconscious processes. In particular, it provides access to appropriately useful resources, thereby solving the relevance problem.

All this takes place under the auspices of contexts: goal contexts, perceptual contexts, conceptual contexts, and/or cultural contexts. Baars uses goal hierarchies, dominant goal contexts, a dominant goal hierarchy, dominant context hierarchies, and lower level context hierarchies. Each context is, itself a coalition of processes. Though contexts are typically unconscious, they strongly influence conscious processes.

Baars postulates that learning results simply from conscious attention, that is, that consciousness is sufficient for learning. There's much more to the theory, including attention, action selection, emotion, voluntary action, metacognition and a sense of self. I think of it as a high level theory of cognition.

"Conscious" Software Agents

A "conscious" software agent is defined to be an autonomous software agent that implements global workspace theory. (No claim of sentience is being made.) I believe that conscious software agents have the potential to play a synergistic role in both cognitive theory and intelligent software. Minds can be viewed as control structures for autonomous agents (Franklin 1995). A theory of mind constrains the design of a "conscious" agent that implements that theory. While a theory is typically abstract and only broadly sketches an architecture, an implemented computational design provides a fully articulated architecture and a complete set of mechanisms. This architecture and set of mechanisms provides a richer, more concrete, and more decisive theory. Moreover, every design decision taken during an implementation furnishes a hypothesis about how human minds work. These hypotheses may motivate experiments with humans and other forms of empirical tests. Conversely, the results of such experiments motivate corresponding modifications of the architecture and mechanisms of the cognitive agent. In this way, the concepts and methodologies of cognitive science and of computer science will work synergistically to enhance our understanding of mechanisms of mind (Franklin 1997).

"Conscious" Mattie

"Conscious" Mattie (CMattie) is a "conscious" clerical software agent (McCauley & Franklin 1998, Ramamurthy et al. 1998, Zhang et al. 1998, Bogner et al. in press) . She composes and emails out weekly seminar announcements, having communicated by email with seminar organizers and announcement recipients in natural language. She maintains her mailing list, reminds organizers who are late with their information, and warns of space and time conflicts. There is no human involvement other than these email messages. CMattie's cognitive modules include perception, learning, action selection, associative memory, "consciousness," emotion and metacognition. Her emotions influence her action selection. Her mechanisms include variants and/or extensions of Maes' behavior nets (1989), Hofstadter and Mitchell's Copycat architecture (1994), Jackson's pandemonium theory (1987), Kanerva's sparse distributed memory (1988), and Holland's classifier systems (Holland 1986) .

IDA

IDA (Intelligent Distribution Agent) is a "conscious" software agent being developed for the US Navy (Franklin et al. 1998). At the end of each sailor's tour of duty, he or she is assigned to a new billet. This assignment process is called distribution. The Navy employs some 200 people, called detailers, full time to effect these new assignments. IDA's task is to facilitate this process, by playing the role of detailer. Designing IDA presents both communication problems, and action selection problems involving constraint satisfaction. She must communicate with sailors via email and in natural language, understanding the content and producing life-like responses. Sometimes she will initiate conversations. She must access a number of databases, again understanding the content. She must see that the Navy's needs are satisfied, for example, the required number of sonar technicians on a destroyer with the required types of training. In doing so she must adhere to some ninety policies. She must hold down moving costs. And, she must cater to the needs and desires of the sailor as well as is possible. This includes negotiating with the sailor via an email correspondence in natural language. Finally, she must write the orders and start them on the way to the sailor. IDA's architecture and mechanisms are largely modeled after those of CMattie, though more complex. In particular, IDA will require improvised language generation where for CMattie scripted language generation sufficed. Also IDA will need deliberative reasoning in the service of action selection, where CMattie was able to do without. Her emotions will be involved in both of these.

"Conscious" Software Architecture and Mechanisms

In both the CMattie and IDA architectures the processors postulated by global workspace theory are implemented by codelets, small pieces of code. These are specialized for some simple task and often play the role of demon waiting for appropriate condition under which to act. The apparatus for producing "consciousness" consists of a coalition manager, a spotlight controller, a broadcast manager, and a collection of attention codelets who recognize novel or problematic situations (Bogner 1999, Bogner et al. in press). Each attention codelet keeps a watchful eye out for some particular situation to occur that might call for "conscious" intervention. Upon encountering such a situation, the appropriate attention codelet will be associated with the small number of codelets that carry the information describing the situation. This association should lead to the collection of this small number of codelets, together with the attention codelet that collected them, becoming a coalition. Codelets also have activations. The attention codelet increases its activation in order that the coalition might compete for "consciousness" if one is formed.

In CMattie and IDA the coalition manager is responsible for forming and tracking coalitions of codelets. Such coalitions are initiated on the basis of the mutual associations between the member codelets. At any given time, one of these coalitions finds it way to "consciousness," chosen by the spotlight controller, who picks the coalition with the highest average activation among its member codelets. Global workspace theory calls for the contents of "consciousness" to be broadcast to each of the codelets. The broadcast manager accomplishes this.

Both CMattie and IDA depend on a behavior net (Maes 1989) for high-level action selection in the service of built-in drives. Each has several distinct drives operating in parallel. These drives vary in urgency as time passes and the environment changes. Behaviors are typically mid-level actions, many depending on several codelets for their execution. A behavior net is composed of behaviors and their various links. A behavior looks very much like a production rule, having preconditions as well as additions and deletions. A behavior is distinguished from a production rule by the presence of an activation, a number indicating some kind of strength level. Each behavior occupies a node in a digraph (directed graph). The three types of links of the digraph are completely determined by the behaviors. If a behavior X will add a proposition b, which is on behavior Y's precondition list, then put a successor link from X to Y. There may be several such propositions resulting in several links between the same nodes. Next, whenever you put in a successor going one way, put a predecessor link going the other. Finally, suppose you have a proposition m on behavior Y's delete list that is also a precondition for behavior X. In such a case, draw a conflictor link from X to Y, which is to be inhibitory rather than excitatory.

As in connectionist models, this digraph spreads activation. The activation comes from activation stored in the behaviors themselves, from the environment, from drives, and from internal states. The environment awards activation to a behavior for each of its true preconditions. The more relevant it is to the current situation, the more activation it's going to receive from the environment. This source of activation tends to make the system opportunistic. Each drive awards activation to every behavior that, by being active, will satisfy that drive. This source of activation tends to make the system goal directed. Certain internal states of the agent can also send activation to the behavior net. This activation, for example, might come from a coalition of codelets responding to a "conscious" broadcast. Finally, activation spreads from behavior to behavior along links. Along successor links, one behavior strengthens those behaviors whose preconditions it can help fulfill by sending them activation. Along predecessor links, one behavior strengthens any other behavior whose add list fulfills one of its own preconditions. A behavior sends inhibition along a conflictor link to any other behavior that can delete one of its true preconditions, thereby weakening it. Every conflictor link is inhibitory. Call a behavior *executable* if all of its preconditions are satisfied. To be acted upon a behavior must be executable, must have activation over threshold, and must have the highest such activation. Behavior nets produce flexible, tunable action selection for these agents.

Action selection via behavior net suffices for CMattie due to her relatively constrained domain. IDA's domain is much more complex, and requires deliberation in the sense of creating possible scenarios, partial plans of actions, and choosing between them. For example, suppose IDA is considering a sailor and several possible jobs, all seemingly suitable. She must construct a scenario for each of these possible billets. In each scenario the sailor leaves his or her current position during a certain time interval, spends a specified length of time on leave, possibly reports to a training facility on a certain date, and arrives at the new billet with in a given time frame. Such scenarios are valued on how well they fit the temporal constraints and on moving and training costs.

Scenarios are composed of scenes. IDA's scenes are organized around events. Each scene may require objects, actors, concepts, relations, and schema represented by frames. They are constructed in a computational workspace corresponding to working memory in humans. We use Barsalou's perceptual symbol systems as a guide (1999). The perceptual/conceptual knowledge base of this agent takes the form of a semantic net with activation called the slipnet. The name is taken from the Copycat architecture that employs a similar construct (Hofstadter & Mitchell 1994). Nodes of the slipnet constitute the agent's perceptual symbols. Pieces of the slipnet containing nodes and links, together with codelets whose task it is to copy the piece to working memory constitute Barsalou's perceptual symbol simulators. These perceptual symbols are used to construct scenes in working memory. The scenes are strung together to form scenarios. The work is done by deliberation codelets. Evaluation of scenarios is also done by codelets.

Deliberation, as in humans, is mediated by the "consciousness" mechanism. Imagine IDA in the context of a behavior stream whose goal is to select a billet for a particular sailor. Perhaps a behavior executes to read appropriate items from the sailor's personnel database record. Then, possibly, comes a behavior to locate the currently available job requisitions. Next might be a behavior that runs information concerning each billet and that sailor through IDA's constraint satisfaction module, producing a small number of candidate billets. Finally a deliberation behavior may be executed that sends deliberation codelets to working memory together with codelets carrying billet information. A particular billet's codelets wins its way into "consciousness." Scenario building codelets respond to the broadcast and begin creating scenes. This scenario building process, again as in humans, has both it's "unconscious" and its "conscious" activities. Eventually scenarios are created and evaluated for each candidate billet and one of them is chosen. Thus we have behavior control via deliberation.

Deliberation is also used in IDA to implement voluntary action in the form of William James' ideomotor theory as prescribed by global workspace theory. Suppose scenarios have been constructed for several of the more suitable jobs. An attention codelet spots one that it likes, possibly due to this codelets predilection for low moving costs. The act of bring these candidate to consciousness serves to propose it. This is James' idea popping into mind. If now other attention codelet brings an objection to conscious or proposes a different job. A codelet assigned the particular task of deciding will conclude, after a suitable time having passed, that the proposed job will be offered and starts the process by which it will be so marked in working memory. Objections and proposals can continue to come to consciousness, but the patience of the deciding codelet dampens as time passes. Several jobs may be chosen with this process.

IDA's language generation module follows the same back and forth to "consciousness" routine. For example, in composing a message offering a sailor a choice of two billets, an attention codelet would bring to "consciousness" the information that this type of message was to be composed and the sailor's name, pay grade and job description. After the "conscious" broadcast and the involvement of the behavior net as described above, a script containing the salutation appropriate to a sailor of that pay grade and job description would be written to the working memory. Another attention codelet would bring this salutation to "consciousness" along with the number of jobs to be offered. The same process would result in an appropriate introductory script being written below the salutation. Continuing in this manner filled in scripts describing the jobs would be written and the message closed. Note that different jobs may require quite different scripts. The appeal to "consciousness" results in some version of a correct script being written.

The mediation by the "consciousness" mechanism, as described in the previous paragraphs is characteristic of IDA. The principle is that she should use "consciousness" whenever a human detailer would be conscious in the same situation. For example, IDA could readily recover all the needed items from a sailor's personnel record unconsciously with a single behavior stream. But, a human detailer would be conscious of each item individually. Hence, according to our principle, so must IDA be "conscious" of each retrieved personnel data item.

These agents are also intended to learn in several different ways. In addition to learning via associative memory as described above, IDA also learns via Hebbian temporal association. Codelets that come to "consciousness" simultaneously increase there associations. The same is true to a lessor extent when they are simply active together. Recall that these associations provide the basis coalition formation. Other forms of learning include chunking, episodic memory, perceptual learning, behavioral learning and metacognitive learning. The chunking manager gathers highly associated coalitions of codelets in to a single "super" codelet in the manner of concept demons from pandemonium theory (Jackson 1987) , or of chunking in SOAR (Laird et al. 1987). IDA's episodic memory is cased based in order to be useful to the perceptual and behavior modules that will learn new concepts (Ramamurthy et al. 1998), and new behaviors (Negatu & Franklin 1999) from interactions with human detailers. For example, CMattie might learn about a new piece of sonar equipment and the behaviors appropriate to it. Metacognitive learning employs fuzzy classifier systems (Valenzuela-Rendon 1991).

Conclusions

Here I hope to have described an architecture capable of implementing many human cognitive functions within the domain of a human information agent. I'd hesitate to claim that this architecture, as is, is fully functioning by human standards. It lacks, for instance, the typical human senses of vision, olfaction, audition, etc. Its contact with the world is only through text. These only the most rudimentary sensory fusion by the agents. They lack selves, and the ability to report internal events. There's much work left to be done.

Acknowledgements

This research was supported in part by ONR grant N00014-98-1-0332, and was produced with essential contributions from the Conscious Software Research

Group including Art Graesser, Satish Ambati, Ashraf Anwar, Myles Bogner, Arpad Kelemen, Ravikumar Kondadadi, Irina Makkaveeva, Lee McCauley, Aregahegn Negatu, Hongjun Song, Allexei Stoliartchouk, Uma Ramamurthy, and Zhaohua Zhang.

References

Allen, J. J. 1995. *Natural Language Understanding.* Redwood City CA: Benjamin/Cummings; Benjamin; Cummings.

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

Baars, B. J. 1997. *In the Theater of Consciousness.* Oxford: Oxford University Press.

Barsalou, L. W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences* 22:577–609.

Bogner, M. 1999. Realizing "consciousness" in software agents. Ph.D. Dissertation. University of Memphis.

Bogner, M., U. Ramamurthy, and S. Franklin. in press. Consciousness" and Conceptual Learning in a Socially Situated Agent. In *Human Cognition and Social Agent Technology,* ed. K. Dautenhahn. Amsterdam: John Benjamins.

Edelman, G. M. 1987. *Neural Darwinism.* New York: Basic Books.

Franklin, S. 1995. *Artificial Minds.* Cambridge MA: MIT Press.

Franklin, S. 1997. Autonomous Agents as Embodied AI. *Cybernetics and Systems* 28:499–520.

Franklin, S., and A. C. Graesser. 1997. Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *Intelligent Agents III.* Berlin: Springer Verlag.

Franklin, S., and A. Graesser. 1999. A Software Agent Model of Consciousness. *Consciousness and Cognition* 8:285–305.

Franklin, S., A. Kelemen, and L. McCauley. 1998. IDA: A Cognitive Agent Architecture. In *IEEE Conf on Systems, Man and Cybernetics.* : IEEE Press.

Hofstadter, D. R., and M. Mitchell. 1994. The Copycat Project: A model of mental fluidity and analogy-making. In *Advances in connectionist and neural computation theory, Vol. 2: logical connections,* ed. K. J. Holyoak, and J. A. Barnden. Norwood N.J.: Ablex.

Holland, J. H. 1986. A Mathematical Framework for Studying Learning in Classifier Systems. *Physica* 22 D:307–317. (Also in Evolution, Games and

Learning. Farmer, J. D., Lapedes, A., Packard, N. H., and Wendroff, B. (eds.). NorthHolland (Amsterdam))

Jackson, J. V. 1987. Idea for a Mind. *Siggart* Newsletter, 181:23–26.

Johnson, V. S. 1999. *Why We Feel: The Science of Human Emotions.* Reading, MA: Perseus Books.

Kanerva, P. 1988. *Sparse Distributed Memory.* Cambridge MA: The MIT Press.

Kolodner, J. 1993. *Case-Based Reasoning.* : Morgan Kaufman.

Laird, E. J., Newell A., and Rosenbloom P. S... 1987. SOAR: An Architecture for General Intelligence. *Artificial Intelligence* 33:1–64.

Maes, P. 1989. How to do the right thing. *Connection Science* 1:291–323.

Maturana, R. H., and F. J. Varela. 1980. *Autopoiesis and Cognition: The Realization of the Living, Dordrecht.* Netherlands: Reidel.

Maturana, H. R. 1975. The Organization of the Living: A Theory of the Living Organization. *International Journal of Man-Machine Studies* 7:313–332.

McCauley, T. L., and S. Franklin. 1998. An Architecture for Emotion. In *AAAI Fall Symposium Emotional and Intelligent: The Tangled Knot of Cognition.* Menlo Park, CA: AAAI Press.

Minsky, M. 1985. *The Society of Mind.* New York: Simon and Schuster.

Negatu, A., and S. Franklin; 1999. Behavioral learning for adaptive software agents. Intelligent Systems: ISCA 5th International Conference; International Society for Computers and Their Applications - ISCA; Denver, Colorado; June 1999.

Ornstein, R. 1986. *Multimind.* Boston: Houghton Mifflin.

Picard, R. 1997. *Affective Computing.* Cambridge MA: The MIT Press.

Ramamurthy, U., S. Franklin, and A. Negatu. 1998. Learning Concepts in Software Agents. In *From animals to animats 5: Proceedings of The Fifth International Conference on Simulation of Adaptive Behavior,* ed. R. Pfeifer, B. Blumberg, J.-A. Meyer , and S. W. Wilson. Cambridge,Mass: MIT Press.

Rolls, E. T. 1999. *The Brain and Emotion.* Oxford: Oxford University Press.

Sloman, A. 1987. Motives Mechanisms Emotions. *Cognition and Emotion* 1:217–234.

Sloman, A. 1998. The ``Semantics" of Evolution: Trajectories and Trade-offs in Design Space and Niche Space. In *Progress in Artificial Intelligence,* ed. H. Coelho. Berlin: Springer.

Valenzuela-Rendon, M. 1991. *The Fuzzy Classifier System: a classifier System for Continuously Varying Variables. In: Proceedings of the Fourth International Conference on Genetic Algorithms.* San Mateo CA: Morgan Kaufmann.

Zhang, Z., D. Dasgupta, and S. Franklin. 1998. Metacognition in Software Agents using Classifier Systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence.* Madison, Wisconsin: MIT Press.

The integration and control of behaviour: Insights from neuroscience and AI

David W. Glasspool
Advanced Computation Laboratory, Imperial Cancer Research Fund, Lincoln's Inn Fields, London, and Institute of Cognitive Neuroscience, University College London dg@acl.icnet.uk

Abstract

Clues to the way behaviour is integrated and controlled in the human mind have emerged from cognitive psychology and neuroscience. The picture which is emerging mirrors solutions (driven primarily by engineering concerns) to similar problems in the rather different domains of mobile robotics and intelligent agents in AI. I review both approaches and argue that the layered architectures which appear in each are formally similar. The higher layer of the psychological theory remains obscure, but it is possible to map its functions to an AI theory of executive control. This allows an outline model of Norman and Shallice's Supervisory Attentional System to be developed.

1 Introduction

Building a functional mind is an ambitious goal. How can the cognitive disciplines - artificial intelligence and cognitive psychology - contribute to such an undertaking? Both psychology and AI are well known for studying small areas of cognition and working with theories of single empirical phenomena. In a full scale cognitive theory two related issues must be addressed, those of *integration* (how are numerous cognitive modules organised into a coherent whole, rather than descending into behavioural chaos?) and *control* (how are the modules to be co-ordinated by an explicit goal?). In this paper I consider a set of theories from AI, neuropsychology and mobile robotics which are concerned with the integration and supervisory control of behaviour. These theories provide converging support for a form of cognitive architecture comprising layered control systems, the lower levels of which contain multiple simple, independent behavioural processes while higher levels are characterised by slower deliberative processes which exercise supervisory control.

A natural question is whether a convergence of this sort can benefit the individual disciplines involved by providing insights from other fields. There are potential benefits for both AI and psychology in this case. In the final part of the paper I describe an example of the way insights from AI, which has tended to concentrate on "higher level" cognitive processes, may benefit psychological theory, which tends not to be so well developed in these areas. Thus an AI theory of agent control can provide a model for higher level supervisory processes in a neuropsychological theory of behaviour control.

2 The organisation of action: A neuropsychological approach.

While a number of theories in psychology have addressed the organisation and control of behaviour, that of Norman and Shallice (1980; 1986) is perhaps the most dominant. The theory is informed both by the slips and lapses made by normal individuals in their everyday behaviour, and by the varieties of breakdown in the control of action exhibited following neurological injury.

2.1 Action lapses and slips

Reason (1984) has studied the slips and lapses made by normal individuals during routine behaviour. Errors in everyday behaviour turn out to be surprisingly common, but can be classified as belonging to a limited set of types. These include errors of place substitution (e.g. putting the kettle, rather than the milk, into the fridge after making coffee), errors of object substitution (e.g. opening a jar of jam, not the coffee jar, when intending to make coffee), errors of omission (e.g. pouring water into a tea pot without boiling it), and errors involving the "capture" of behaviour by a different routine (such as going upstairs to get changed but getting into bed). Interestingly Reason finds that the situations in which such slips and lapses occur share two properties in common: The action being performed is well-learned and routine, and attention is distracted, either by preoccupation or by some external event.

There are two points of interest here. Firstly it is clear that we can perform a wide range of often complex habitual actions without concentrating on them - the control of well-learned action can become automatic. Secondly, when we allow such behaviour to proceed without our conscious control it is susceptible to a specific range of characteristic errors. These observations provide one class of data which psychological theories of action control must address. Another important class of data is provided by the effects of neurological damage.

2.2 Neurological impairment of behaviour control

The breakdown of cognitive systems following neurological damage constitutes an important source of constraint on psychological theory. Cooper (2000) reviews a range of problems with the control of action which mainly follow damage to areas of prefrontal cortex. Here I briefly mention three syndromes of particular interest.

Patients with action disorganisation syndrome (ADS, Schwartz et al. 1991, Humphreys & Forde, 1998) make errors which are similar in type to those of normal individuals - errors in the sequencing of actions, the omission or insertion of actions, or the substitution of place or object. However their errors are far more frequent. For example patient HH of Schwartz et al. (1991) made 97 errors during 28 test sessions in which he made a cup of coffee.

Utilisation behaviour (Lhermitte, 1983) can be characterised as weakening of intentional control of action, so that irrelevant responses suggested by the environment may take control of behaviour. A neurological patient exhibiting utilisation behaviour may pick up and perform actions with items lying around on a table, for example, which are appropriate to the items but not relevant to the task in hand.

Shallice and Burgess (1991) report patients with "strategy application disorder" who are able to carry out individual tasks but have difficulty co-ordinating a number of simultaneous task demands. Such patients for example may be able to carry out individual food preparation tasks but are unable to plan and cook a meal. Their deficit appears to be in the ability to schedule multiple tasks over an extended period.

2.3 The Norman and Shallice framework for behaviour control

The challenge for a psychological account of the integration and control of behaviour is to explain data of the type outlined above. Norman and Shallice (1980; 1986) interpret the data as implying that two distinct systems operate to control the range of behaviour typically studied by psychologists. The systems are arranged in a layered manner as shown in Figure 1 (a). Over-learned or habitual action is held to be controlled by a set of *schemas* competing within a contention scheduling (CS) system for control of the motor system, while willed or attentional control of action is achieved by a supervisory attentional system (SAS) which can influence the CS system but has no direct access to motor control.

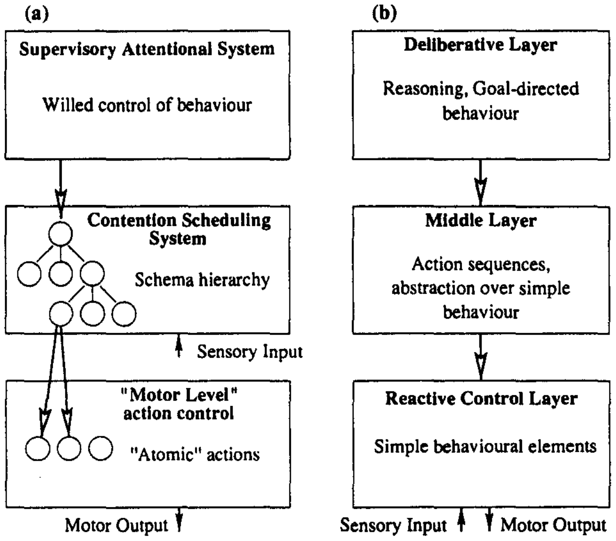

(a)
Supervisory Attentional System
Willed control of behaviour
Contention Scheduling System
Schema hierarchy
Sensory Input
"Motor Level" action control
"Atomic" actions
Motor Output
(b)
Deliberative Layer
Reasoning, Goal-directed behaviour
Middle Layer
Action sequences, abstraction over simple behaviour
Reactive Control Layer
Simple behavioural elements
Sensory Input
Motor Output

Figure 1: (a) Norman and Shallice's (1986) framework for action control augmented with Cooper and Shallice's (in press) distinction between cognitive and motor level action. (b) The three-layer architecture of Gat (1998) and colleagues.

Cooper and Shallice (in press) provide a number of arguments for distinguishing, on grounds of psychological data, two sub-levels of low-level behaviour. The lower sub-level, "motor" behaviour, comprises the individual motor commands required to carry out a simple action (extending and retracting individual muscle groups to grasp an item, for example). The higher sub-level, the "cognitive" level, operates with actions at the lowest level to which they are referred in everyday language - grasping, reaching etc. Norman and Shallice's CS component applies to cognitive level actions, which abstract over motor level actions. The theory does not directly address operations at the motor level.

The contention scheduling system comprises a hierarchy of schemas, defined as discrete actions or structures organising sets of actions or lower-level schemas. The schema hierarchy terminates in a set of "cognitive level" actions which are held to be carried out directly by motor systems. Actions at this level might include, for example, "pick up an item", "unscrew", or "stir". Higher level schemas might include "open jar", which would organise the actions of picking up, unscrewing a lid, and putting down. At a higher level still a "make coffee" schema might exist.

Schemas are connected in an interactive-activation network. They are activated from the top down by their parent schemas or by control from the SAS, and from the bottom up by input from the environment. They compete for execution on the basis of their activation level. A schema is triggered when its activation level is higher than any other schema and higher than a trigger threshold. A triggered schema feeds activation forward to its child schemas, and is inhibited after its goal has been achieved. Top-down activation can exert detailed control over behaviour or it can simply be used to specify goals, by activating high-level schemas. Such schemas may provide multiple ways for a goal to be achieved - coffee can be supplied in a jar or a packet, for example, so a schema for adding coffee to a mug can be indifferent to the particular lower level behaviour required to achieve its goal. Whichever suitable sub-schema best fits the current configuration of the environment will be selected.

Cooper and Shallice (in press) have simulated the CS system in detail. With a certain amount of background noise in the system, and a reduction in top-down input, the system makes occasional errors analogous to those made by normal individuals, when the wrong schema or sub-schema is triggered. By varying the parameters of the model - in particular the levels of top-down influence and environmental influence, utilisation behaviour and ADS can be simulated, as well as a number of other neuropsychological disorders of action control.

Just as important to the Norman and Shallice account of behaviour control is the SAS, which is held to take control of behaviour in non-routine situations (ie those where no appropriate well-learned schema exists) and in situations where performance is critical. The SAS exerts control by directly activating individual low-level actions, or by causing the selection of an existing schema which would not otherwise be selected in that situation. Internally, however, the SAS is poorly specified. Based largely on neuropsychological evidence but partially guided by *a priori* reasoning about the types of processes which must be involved in supervisory processing, Shallice and Burgess (1996) set out an outline of the processes involved in the SAS and their relationships during supervisory processing. They characterise the functioning of the SAS as centrally involving the construction and implementation of a temporary new schema, which can control lower level CS schemas so as to provide a procedure for dealing effectively with a novel situation.

Shallice and Burgess' characterisation of the SAS as modular, and their preliminary functional decomposition, provide a useful starting point for neuropsychological theory. However the picture remains unclear, with many processes under-specified. This is largely due to the difficulty of obtaining clear empirical data on such high-level processes. We return to the specification of the SAS later. For now however we can note that it is concerned with problem solving and planning, and delegates the control of routine behaviour to the CS system as long as things are running smoothly.

The Norman and Shallice theory provides a framework for the control of willed and automatic behaviour based on psychological and neuropsychological evidence. I now turn to an equivalent problem in artificial intelligence - the control of behaviour in autonomous robots.

3 The organisation of action in mobile robotics

Mobile robotics has long been seen as an important area for artificial intelligence research. It is an area where all aspects of an agent's behaviour and its interaction with its internal and external environment must be taken into account. Theories are forced to address, to some extent at least, the entire cognitive system from sensory input to motor output, and the interaction of the agent with its environment.

Early AI robotics projects (e.g. "Shakey", Nilsson 1984; the CART, Moravec, 1982) employed architectures centering on classical planning systems. Such systems typically involve three sequential steps in their control architectures: sensing, planning and acting. In the first step sensory information (e.g. from a video camera) is analysed and used to form a map of the robot's environment. In the second step a search-based planning system is applied to the map to find the most appropriate plan of actions to be followed in order to achieve a goal. Once a plan has been generated the robot can make a move. Such systems are often known as sense-plan-act (SPA) architectures.

There are a number of well-known problems with this approach. It requires search over a large state-space, leading to slow, resource-hungry operation. The plan which is generated is critically dependent on the reliability of the sensors and on the environment remaining static while the plan is formulated. Even with improvements in computing hardware and planning techniques robots based on this paradigm tend to remain slow, cumbersome and fragile in their operation.

In the mid 1980s Brooks developed an alternative approach to robot control in response to these problems, sometimes termed reactive control (or "reactive planning", Brooks 1991). This represents a break from the sense-plan-act cycle. Brook's paradigm largely does away with a central representation of the world and uses many simple, high-speed (reactive) processes coupling simple sensory systems directly to action, operating in a highly parallel manner. These reactive processes implement small, circumscribed elements of behaviour, and are usually referred to simply as "behaviours". The direct coupling of input to output and decomposition of behaviour into many simple, environmentally-driven "behaviours" allows small, fast, robust and flexible robot control systems to be built.

Rapid theoretical development followed Brook's initial work. It soon became apparent that, in its pure form, Brooks' reactive behaviour paradigm becomes difficult to program as more complex behaviour patterns are attempted. In practical applications the lack of any ability to carry out high-level planning and problem solving was also a concern. Gat and colleagues (Gat, 1998) have been in the vanguard of a second wave of development aimed at formalising reactive agent control systems to make them more robust and scalable. Much of this work centres on the idea that three distinct layers of control are required for a large-scale practical agent: a rapid but simple reactive low-level control system, an intermediate system capable of stringing together sequences of simple actions into useful behavioural elements, and a slow "deliberative" high level system capable of carrying out more complex planning and reasoning. Such schemes have been termed three-layer architectures (TLAs, Gat 1998) (Figure 1, b).

The lowest level in a TLA provides the responsive, flexible and robust low-level control of behaviour characteristic of Brooks' reactive approach. The top level provides a more traditional AI planning and problem-solving capability, allowing the robot's behaviour to be guided by long term, abstract goals. The middle layer interfaces between the two. It provides abstractions over lower level behaviours in two ways - by constructing more powerful behavioural elements through assembling sequences of simple behaviours, and by providing higher level goals which may be achieved by different lower level actions depending on prevailing circumstances. The top level system can interact with the robot through relatively abstract commands and need not specify every detail of the actions needed to implement its goals.

4 Converging architectures?

The Norman and Shallice framework and the TLA paradigm address similar issues of control and integration of an agent's behaviour in two rather different domains. While the original Norman and Shallice theory speaks to only two layers of control - CS and SAS - the inclusion of Cooper and Shallice's "motor" action level yields a three-layer framework. The correspondence with the TLA is striking (Figure 1). Might the resemblance simply be superficial, though? We need to compare the way the layers are specified in each approach.

Shallice and Burgess describe the SAS as corresponding to frontal-lobe processes "critically involved in coping with novel situations as opposed to routine ones" (1996, p.1406). They specify its functions in terms of goal-setting, problem solving and schema generation (planning). Gat (1998) describes the topmost TLA system as "the locus of time-consuming computations. Usually this means such things as planning and other exponential search-based algorithms [...] It can produce plans for the [middle layer] to implement, or it can respond to specific queries from the [middle layer]". In other words the main functions are generating new plans of action and dealing with situations for which no pre-existing procedure exists in lower levels, i.e. novel situations. Despite the language differences - an inevitable consequence of comparison across disciplines - the two architectures apparently ascribe essentially the same functions to their highest level systems.

Turning to the lowest level of behaviour control, on Cooper and Shallice's (in press) account this corresponds to "motor level" actions. These operations are the preserve of motor systems and are not susceptible to the types of errors typically made at the "cognitive" level. On the Norman & Shallice / Cooper & Shallice framework the distinction between the lowest (motor) level and middle (CS) level is well defined. It is not clear that the corresponding distinction in the TLA approach is well defined, however. Gat (1998) describes the processes at the lowest TLA level as "designed to produce simple primitive behaviours that can be composed to produce more complex task-achieving behaviour". The composition of simple behaviours into complex behaviour is a function of the middle layer. It is not entirely clear at what point a simple behaviour becomes a complex one (although Gat does give a number of guidelines for the type of behaviour to be considered simple, including keeping internal state to a minimum and using only input-output transfer functions which are continuous with respect to internal state). If the idea were simply that actions which are, from the point of view of higher level systems, atomic should be included this level would correspond well with Cooper and Shallice's motor level. However the notion of reactive control - tight sensory-to-motor coupling - is an important part of the TLA definition of this layer. The triggering of action by environmental input is not prominent in Cooper and Shallice's characterisation (although reflex and sensory-motor feedback certainly play an important part in low-level human motor control). This type of control is however certainly part of the definition of CS. Cooper and Glasspool (in submission), for example, treat the environmental triggering conditions of schemas in CS as "affordances" for action, priming appropriate behaviour in response to learned environmental configurations. It is thus possible that the lowest level layer in the TLA account corresponds to a combination of the motor layer and the lowest level action representations in CS. Higher order schemas in CS would then correspond to the middle TLA layer.

In the TLA account, a primary function of the middle layer is to organise primitive behaviours into behaviour sequences which perform two functions: they form a more compact and convenient representation of behaviour for use by higher level processes (i.e. sequences of behaviour which are often needed are "chunked" together), and they provide abstraction - alternative means may be specified for achieving a goal, providing low-level flexibility and avoiding the need to specify behaviour in detail. Both of these functions are central to the Norman and Shallice CS system. Schemas represent well-learned fragments of behaviour and provide a goal-based representation - sub-schemas for achieving the same goal compete to service a higher-order schema's requirements. Functionally, the CS corresponds well to the TLA middle layer.

In this connection it is important to note an early attempt to overcome some of the problems of "pure" reactive robotic control by Maes (1989). Maes' scheme has a range of alternative behaviours (specified at a level typical of the TLA "middle layer") competing for control of resources (robot effectors) in an interactive activation network under the influence of environmental input. The similarities with Contention Scheduling are striking, especially given the very different provenance of the theories. The approach has not been followed up, apparently because of a view that in real-world cases robot control systems can be made simple enough that flexible, on-line resource allocation and conflict resolution are not necessary. That this appears to be a primary function of intermediate-level behaviour control in humans suggests that this view may be challenged as robotic systems are scaled up to more complex tasks.

It thus appears that the similarity between TLAs and the SAS/CS framework is more than superficial and may represent a true convergence of theory in two distinct areas. Whether this is the case would be clearer with a more detailed specification of the Norman and Shallice framework. The CS component is well specified and has been modelled in detail by Cooper and Shallice (in press). The motor level and the SAS are less clearly specified. The SAS in particular is only characterised in outline by Shallice and Burgess (1996). However, an implementation of the SAS, even if only in outline, would provide a valuable first step in fully formalising the theory as well as enabling a number of issues concerning the interface between SAS and CS to be addressed. In the remainder of this paper I therefore describe a first step towards a computational model of the SAS.

5 Modelling the SAS

The shadowy nature of the SAS is testament to the difficulty of "reverse engineering" processes of such scope and complexity in human psychology. However, while the SAS is a construct posed at an unusually high level for psychological theory, it does address processes at the same general level as many theories in AI. This may allow psychological theory to benefit from the alternative perspective of AI, with its greater emphasis on engineering intelligent systems from first principles. Shallice and Burgess (1996) identify three stages in the operation of the SAS in its typical role of reacting to an unanticipated situation:

1. The construction of a temporary new schema. This is held to involve a problem orientation phase during which goals are set, followed by the generation of a candidate schema for achieving these goals.

2. The implementation of the temporary schema. This requires sequential activation of existing schemas in CS corresponding to its component actions.

3. The monitoring of schema execution. Since the situation and the temporary schema are both novel processing must be monitored to ensure that the schema is effective.

The domino model of Fox and colleagues (Das, Fox, Elsdon & Hammond, 1997; see also Fox and Cooper, this symposium) provides a framework for processes of goal-setting, problem solving and plan execution which gives a promising initial fit to Shallice and Burgess's outline. It specifies seven types of process operating on six types of information. The domino framework is shown in Figure 2 (broken lines). Starting from a database of beliefs about its environment the agent raises goals in response to events requiring action. Such goals lead to problem solving in order to find candidate solutions. Alternative solutions are assessed and one is adopted, leading to new beliefs and possibly to the implementation of a plan of action, which is decomposed into individual actions in the world. The processes are similar to those specified by Shallice and Burgess: goal setting, solution generation and evaluation, decision making, planning, acting and monitoring the effects of action. A set of well understood and well specified formal semantics can be associated with the framework to render it computationally implementable. The domino thus provides an appropriate starting point for an SAS model. Figure 2 shows that the processes identified by Shallice and Burgess (1996) can be mapped cleanly onto the domino framework. The "candidate solution generation" process of the domino framework corresponds to the generation of a "strategy" in SAS - a generalised plan of action which is subsequently implemented as a concrete schema for execution by the CS system.

5.1 Architecture and operation

For the purposes of modelling a target task is required. A standard test of frontal lobe (and *ex hypothesi* of SAS) function in neuropsychology is the Wisconsin card-sorting test (WCST). The subject is given a set of cards which vary in the number, shape and colour of the symbols they show (thus a card might show two green squares, or four red triangles). The experimenter lays out four "stimulus" cards, and the subject is asked to sort the cards into piles corresponding to these, but they are not told the criterion for sorting. They might sort cards by the number of symbols, their colour or their shape. After each card is placed the experimenter indicates whether it was correctly sorted. Once the subject has worked out the sorting criterion the experimenter is using they are allowed to place ten cards correctly, then the experimenter changes to another sorting criterion without warning. Neurologically intact individuals typically catch on to the procedure quickly and make few errors, these being immediately after the change of criterion. Patients with frontal lobe damage make many errors, typically involving the inability to discover the sorting strategy or inability to change strategies despite repeated negative feedback.

Sorting objects according to their features is the type of well-learned behaviour we would expect to find as a high-level schema in CS. The CS/SAS model would most

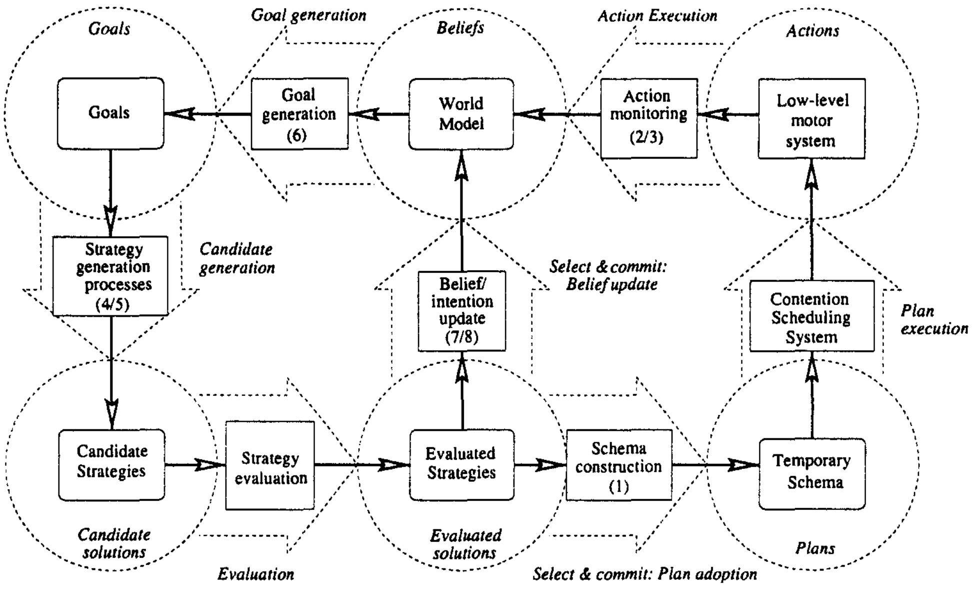


Figure 2: The SAS outline of Shallice and Burgess (1996) mapped on to the Domino framework of Das et al. (1997) (broken lines). Numbers in brackets refer to processes as identified by Shallice and Burgess.

straightforwardly address the WCST on the basis that SAS is involved in initial generation of a sorting strategy and configuration of CS, which would then carry out that strategy with subsequent cards unless negative feedback was received, when SAS is again required to generate an alternative strategy. Figure 3 shows a minimal implementation of the system of Figure 2 in the COGENT computational modelling environment which allows the boxes in such "box and arrow" diagrams to be fleshed out with computational specifications so that the model may be executed.

Bearing in mind that at this stage the requirement is simply for an outline model to demonstrate the principle of an SAS implementation, the implementation of Figure 3 is simplified to include only the essential elements of Figure 2. Following Figure 3 in a clockwise direction operation is as follows: "Current beliefs" maintains information from the environment provided by sensory processes. A "Novelty detection" process triggers the generation of a new goal in response to an unexpected situation, which may be the result of novel circumstances or of the failure of an automatised behaviour in CS. The presence of a goal triggers strategy generation processes. A number of such processes may operate in parallel on the problem posed by the goal, potentially yielding more than one candidate solution. A solution evaluation process provides a means of ranking these candidates, yielding the fourth domino "dot", Evaluated Strategies. At this point the highest ranked candidate is selected for implementation. The "current beliefs" are updated to reflect the candidate strategy. Simultaneously, the strategy is enacted via the CS system. This may simply require the activation of an existing CS schema or may involve the construction and implementation of a new temporary schema. A single process (Schema Implementation) is assumed to be responsible for either, resulting in a temporary schema specification which sends activation to existing CS schemas.

Shallice and Burgess suggest a number of procedures for strategy generation in response to a goal, the simplest of which is "spontaneous schema generation" - the propensity of a suitable strategy to simply come to mind in response to a simple problem. In the current implementation a process of this type is simulated by a rule in the "strategy generation" process which may be paraphrased as: If the goal is to sort an item into a category, and the item has distinguishable features, the item may be sorted according to one of those features. Cards are defined as having the features symbol, number and colour, so this rule will always generate three corresponding sorting strategies. The "strategy evaluation" process ranks strategies according to two rules: Strategies which have recently been attempted are ranked lower, and strategies which have recently proved successful are also ranked lower. A strategy which has recently been attempted and has been successful will thus be ranked lowest of all. This simple scheme leads to appropriate strategy-testing behaviour during the WCST task.

The Contention Scheduling system is simulated in the current model by a simple set of processes; a full computational simulation is available which could be used for more detailed modelling (Cooper & Shallice, in press). A single well-learned schema ("match_to_feature") is assumed to be present for placing a held item next to a stimulus matching on a specified feature. This schema may be activated by the SAS simulation along with a token representing the feature to be matched (colour, shape or number).

Performance of the WCST task starts with a request from an external "experimenter" process to sort a card.

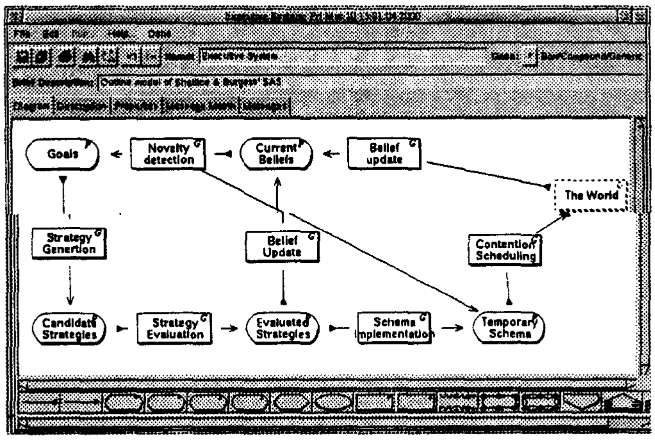


Figure 3: An outline implementation of the Shallice and Burgess SAS in the COGENT modelling system. Rounded boxes are buffers, square boxes are processes. "The world" is an external world representation.

This is placed in "current beliefs" and is treated as a novel event. A goal is thus set to serve this request. This triggers strategy generation which produces three candidate strategies, sort by colour, shape or number. Initially all are equally ranked so one is selected at random for execution. This leads to update of beliefs (with the new current strategy) and to execution of the strategy, which involves activation of the "match_to_feature" schema along with the corresponding feature token in the CS simulation. This schema executes in CS causing the current card to be matched according to the chosen feature. If this action receives positive feedback from the experimenter the SAS takes no further action - as further cards are produced by the experimenter the "match_to_feature" schema remains active and immediately responds by sorting them appropriately. If the experimenter gives negative feedback (which may occur immediately if the wrong sorting strategy has been attempted first, or may occur after a number of correct sorts when the experimenter changes the sorting criterion) the SAS treats this as a novel situation and again raises a goal to find a sorting strategy. Recently tried and recently successful strategies are both ranked lower than untried strategies ensuring a that successful new strategy is rapidly found.

The simulation raises a problem at this point, however. While the SAS simulation is determining a new strategy the CS simulation still has the old strategy active and proceeds to sort the next card despite the negative feedback. Evidently an additional control signal is required to halt automatic behaviour in CS when unexpected feedback is received. Intuitively this seems reasonable: animals have a "startle" reflex which achieves much this result in situations where the habitual response needs to be suppressed. A connection is accordingly added to the simulation (between "novelty detection" and the temporary schema in Figure 3) which removes the current temporary schema when triggered. This in turn removes activation input from the currently active CS schemas and halts automatic behaviour. Table 1 shows sample output from a short run of the WCST simulation.

Table 1: Sample output from a short run of the WCST simulation. The experimenter's criterion is initially to sort by shape, but changes to sort by colour after three correct responses.

Card to sort	Model's response	Feedback
4 blue squares	place with 4s	wrong
2 green triangles	place with triangles	correct
1 red square	place with squares	correct
3 blue circles	place with circles	correct
2 green circles	place with circles	wrong
1 red triangle	place with reds	correct
2 blue squares	place with blues	correct

5.2 Discussion

While the model described here is certainly highly simplified and just as certainly incomplete, it represents a first step towards a psychologically plausible simulation of major aspects of the SAS. The current simulation is not detailed enough to allow very specific claims to be made about the origin of errors in the WCST following frontal-lobe damage, but some general points can be raised. The best known error type, perseverative responding (i.e. failure to adjust to a new sorting strategy when the experimenter changes the sorting criterion) may implicate a number of systems. For example, negative feedback may fail to result in the generation of a goal to change behaviour; candidate strategies may not be correctly weighted, so that the previously successful strategy is chosen again despite having been recently used and having elicited negative feedback; or the process of de-selecting the current schema in CS may be defective. Perseverative behaviour can be simulated in the model in any of these ways and a more detailed simulation, including a full simulation of contention scheduling, may provide a better basis for disambiguating these possibilities.

A more general benefit of an SAS simulation is the possibility of investigating the interface between SAS and CS. Learning is one important target for investigation. The CS system is held to acquire new schemas as a result of repeated application of the same strategy by SAS in similar situations. Once a schema has been acquired the SAS is able to delegate operation to it without having to explicitly control behaviour. A number of processes are implicated in this SAS-to-CS transfer which cannot be studied without adequate characterisations of the two systems.

Another aspect of the interaction between SAS and CS is the need to remove the temporary schema (and possibly also deselect CS schemas) in response to novelty. Interestingly such behaviour is also found in robot control systems where a sufficiently powerful top-level executive system is present. For example an autonomous spacecraft control system demonstrated recently by NASA (Muscettola et al. 1998) includes a process which puts the spacecraft into a "standby" mode - suspending routine operations - when an anomalous event occurs. Operation resumes when the anomaly has been analysed by executive systems and a new plan of action generated to deal with it. The need to add this behaviour to the model illustrates the advantage of simulation in the analysis of large-scale agent models. The interactions of multiple systems controlling behaviour with each other, with the agent as a whole and with its environment can be difficult to analyse in the abstract.

6 Conclusions

I have argued that architectures for the integration and control of behaviour which have emerged from the study of neuropsychological data and from essentially engineering research into the efficient control of mobile robots are formally similar. While competing positions exist in both fields the apparent convergence of independent work in different domains indicates that this class of mechanism is worth investigation as a candidate architecture for a functional model of mind. Within cognitive psychology a major problem is the obscurity of higher-level processes. I have suggested that theories in AI, which are typically more focussed on higher cognitive functions, may point the way to appropriate decompositions of such opaque processes, and I have offered a preliminary model of the Norman and Shallice SAS as an example.

Theories have been constructed in AI and in cognitive psychology which address the same types of cognitive process, and both disciplines have made great progress in recent years in adding detail to these theories. It seems that both have now reached a level where we can expect each to begin providing useful insights for the other. A dialogue between AI and neuroscience on the problem of the control and integration of behaviour should benefit both fields. Approaches from AI and robotics may shed light on the structure of obscure higher processes in psychology. In turn the increasingly detailed picture of human executive function emerging from neuropsychology can provide a rich context for theories of behaviour integration and control in AI.

Acknowledgements

I am most grateful to Richard Cooper, John Fox, Tim Shallice and Heather Rhodes for numerous discussions on the material in this paper.

References

R. A. Brooks. Intelligence without representation. *Artificial Intelligence 47*, 139-160. 1991.

R. Cooper. The control of routine action: modelling normal and impaired functioning. To appear in G. Houghton (ed.) *Connectionist Modelling in Psychology*. Psychology Press. 2000.

R. Cooper & D. W. Glasspool. *Learning to act.* In submission.

R. Cooper & T. Shallice. Contention Scheduling and the control of routine activities. *Cognitive Neuropsychology*. In press.

S. K. Das, J. Fox, D. Elsdon & P. Hammond A Flexible architecture for autonomous agents. *Journal of Experimental and Theoretical Artificial Intelligence*, 9, 407-440. 1997.

E. Gat. On three layer architectures. In D. Kortenkamp, R. P. Bonnasso and R. Murphey, eds. *Artificial Intelligence and Mobile Robots.* AAAI Press. 1998.

G. W. Humphreys & E. M. E. Forde. Disordered action schema and action disorganisation syndrome. *Cognitive Neuropsychology 15*, 771-811. 1998.

F. Lhermitte. Utilisation behaviour and its relation to lesions of the frontal lobes. *Brain*, 106, 237-255. 1983.

P. Maes. How to do the right thing. *Connection Science*, 1, 291-323. 1989.

H. P. Moravec. The Stanford Cart and the CMU Rover. *Proceedings of the IEEE*, 71(7), 872-884. 1982.

N. Muscettola, P. Nayak, B. Pell & B. Williams. Remote Agent: to boldly go where no AI system has gone before. *Artificial Intelligence* 103(1-2):5-48. 1998.

N. J. Nilsson (Ed.). *Shakey the robot.* SRI AI Center technical note 323. 1984.

D. A. Norman & T. Shallice. *Attention to action: Willed and automatic control of behaviour. Center for Human Information Processing* (Technical Report No. 99). University of California, San Diego. 1980.

D. A. Norman & T. Shallice. Attention to action: Willed and automatic control of behaviour. Reprinted in revised form in R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.) 1986. *Consciousness and self-regulation, Vol. 4* (pp. 1-18). New York: Plenum Press. 1986.

J. T. Reason. Lapses of attention in everyday life. In W. Parasuraman and R. Davies (eds) *Varieties of Attention*, pp. 515-549. Orlando, FL: Academic Press. 1984.

M. F. Schwartz, E. S. Reed, M. W. Montgomery, C. Palmer & N. H. Mayer. The quantitative description of action disorganisation after brain damage: a case study. *Cognitive Neuropsychology*, 8, 381-414. 1991.

T. Shallice & P. Burgess. The domain of supervisory processes and temporal organization of behaviour. *Philosophical Transactions of the Royal Society of London B.* 351, 1405-1412. 1996.

T. Shallice & P. Burgess. Deficits in strategy application following frontal lobe lesions. *Brain*, 114, 727-741. 1991.

An Artificial Mind via Cognitive Modular Neural Architecture

Pentti O. A. Haikonen

Nokia Research Center
P.O. Box 407
FIN-00045 Nokia Group, Finland
pentti.haikonen@nokia.com

Abstract

The author proposes that an artificial mind should be able to duplicate the processes of the human mind, i.e. inner imagery, inner speech, sensations, the cognitive functions like introspection, perception, attention, match, mismatch and novelty detection, learning, memory, reasoning, planning, emotions and motivation, perhaps even consciousness. Furthermore, the author proposes that a cognitive system's ability to perceive and report its inner imagery as such should be taken as a test for machine self–consciousness.

An artificial cognitive system based on modular neural architecture is presented here. This non-numeric system utilizes distributed signal representation, sensory preprocessing into feature signals, processing of information associatively with meaning and significance and a modular reentrant architecture that allows the establishment of inner imagery and speech as well as introspection. Match, mismatch and novelty signals are derived from neuron-level signal relations and they are used to effect sensory and inner attention. Pleasure/displeasure conditions are also modelled and they contribute to the reactive state of the system.

This cognitive system is simulated by a PC with real digital camera and text input. The simulated system architecture consists of perception/response reentrant loop modules for linguistic, visual and gaze direction subsystems as well as subsystem modules for pleasure/displeasure and match/mismatch evaluation. The perception/response reentrant loop realizes sensory perception primed by prediction and inner evocations, introspective perception and the establishment and the grounding of meaning for inner imagery and inner words. Additionally the reentrant loop acts as a reverberating short-term working memory. In-loop associative neuron groups facilitate associative cross-connections to other modules. Learning and long-term memory are realized via synaptic strength modifications.

The system can learn to recognize figures, learn the meaning of concrete words by ostension and via correlation, learn certain abstract words and rudimentary syntax by examples, learn to recognize new figures by verbal description, learn temporal sequences and predict their continuation, detect affirmation and contradiction, deduct the properties of a given object from evoked inner imagery. Learning is inductive and fast, only few repetitions are needed.

This system has several features that are commonly attributed to consciousness: It is perceptive, it has inner imagery and inner speech; it is introspective, the inner workings are perceived by the system via reentry to perception process; there is attention and short-term memory. However, at this moment the system does neither have a body reference for self-concept nor episodic memory capacity for personal history, these will have to be added later.

1. Introduction

The human mind is characterized by the flow of inner imagery, inner speech, sensations, emotional moods and the awareness of these; consciousness. The human mind is imaginative, creative and intelligent, it can produce correct responses from minimal cues. The human mind possesses intentionality; it operates with meanings and significance, it understands what it is doing. The human mind seems to unify effortlessly past experience, present multisensory information, the expected and desired future, the needs, drives and goals, "the own will" and the emotional states, moods, arising from the interaction of the above. While doing this the human mind seems not to be plagued by the combinatorial explosion.

Obviously the above mentioned qualities would be very useful to any robot, agent or personal electronic assistant.

Would it be possible to reproduce these qualities artificially, would it be possible to create an artificial mind, a thinking machine? How should we proceed towards the design of this kind of machine?

Traditionally two different approaches have existed here.

The symbolic, rule-based artificial intelligence (AI) tries to achieve this goal through programmed processes and functions. Ultimate success has been elusive and strong criticism has arisen (E.g. Devlin 1997, Lenat 1995, Searle 1984, Omar 1994).

The connectionist or artificial neural network (ANN) approach was originally inspired by the biological neuron. Ultimate success has again been elusive and nowadays the research in this field has largely been reduced into the production of isolated functions like pattern recognition or classification and artificial neural networks can be seen mainly as another style of numeric computation. However, at the end of last century there were some bold attempts towards actual cognitive neural systems (Aleksander 1996, Trehub 1991, Valiant 1994).

The problem with AI and ANN approaches has been that the programs and computations do not really understand what they are doing. Meaning and significance are not really involved in the process.

On the other hand human cognition seems to operate with meaning and understanding. Therefore the author proposes that human cognition should be taken as a model for thinking machines and furthermore, a complete system with rich interactions and reactions should be considered instead of rather artificial modelling of isolated cognitive functions with the usual arbitrary labelling of said computations with a cognitive name.

Cognitive psychology has identified basic cognitive processes like perception, attention, learning, deduction, planning, motivation, etc. (Aschcraft 1998, Nairne 1997). Cognitive brain research has also been advancing and the functions of various parts of the brain are being modelled and their possible relationship to cognition and consciousness are being evaluated (E.g. Taylor 1999). Consequently, the author presents here another approach towards thinking machines, based loosely on ideas about the architecture of the brain and on the emulation of cognitive functions by modular non-numeric associative neural networks (Haikonen 1999b, 1998a, 1998b).

2. The Cognitive Approach

The cognitive approach involves the design of a system that is able to process information with meaning in the style of human cognition. This style would mean the reproduction of the flow of "inner speech", inner imagery, the basic cognitive processes like perception, attention, learning, deduction, planning, motivation, etc. and ultimately the awareness of these.

Information processing by meaning and significance involves the understanding of the subject matter. Let's consider the following cases of understanding:

-Scene (image) understanding
-Episode understanding
-Story understanding (narratives, books, movies)

What would constitute understanding in these cases? Obviously *neither* tape recorder type storage and playback *nor* mapping one set of symbols into others *but* the ability to:

-Answer questions about the subject like:
    -what is where
    -what is happening
    -who is doing what to whom, etc.
-Paraphrase; describe with own words
-Detect contradictions
-Predict what happens next, what is possible
-Give reasons for present situation
-Evaluate significance, good/bad/urgent

Accordingly the system requirements for understanding would be:

-Recognition of components; objects, sounds, words, etc.
-Detection of their relationships; spatial, temporal
-Learning and evocation of relevant associations as the story evolves; meanings, context, background
-Prediction
-Deduction, reasoning
-Match/mismatch detection, contradiction detection
-Significance evaluation good/bad, urgent
-Suitable working and episodic temporary memories
-Suitable long term memories
-Avoidance of combinatorial explosion by sensory and inner attention guided by significance etc.
-Information representation and manipulation methods that allow these operations fluently.

"Understanding" involves the evocation on the relevant meanings among all the possible meanings for the subject representations; purpose, relations, names, etc. These meanings and significance are acquired by the system via learning.

Distributed signal representation (Hinton et al. 1990) and non-numeric associative processing with necessary controlling mechanisms are seen here as the methods that allow the realization of the above.

3. The Associative Neuron

Processing with distributed signal representations calls for the ability to connect representations to each other so that one representation can be evoked by another. Individual signals are the basic components of distributed signal representation. Therefore the processing with distributed signal representations involves operations with individual signals. A signal derives its meaning from the point of origin and can be either on or off. Therefore the basic signal processing unit, the neuron, shall switch a signal on or off while preserving the point-of-origin-path and learn when to do the switching. The point-of-origin-path can be preserved if the neuron is configured so that the signal passes through it.

The author has designed a non-numeric associative neuron along the above principles (Haikonen 1999a). This neuron preserves the meaning of signals and thus allows consistent internal representations in suitable network architectures.

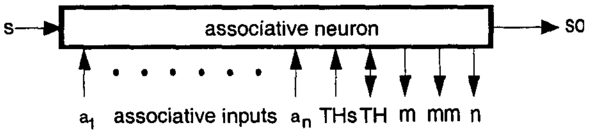


Fig. 3.1. The associative neuron

The associative neuron has one main signal input *s*, a number of associative input signals a_i, a synaptic learning fixation threshold control input *THs*, a neuron output bi-directional threshold control *TH* and the output signals *so*, *m*, *mm* and *n*. The point-of-origin meanings of the *s* and *so* signals are the same. *m*, *mm* and *n* signals represent match, mismatch and novelty conditions. The number of associative input signals a_i is not limited or need to be fixed as their synaptic weights are not adjusted against each other. All signals are assumed to have values between zero and one only.

Synaptic learning is correlative and depends on repeated *s* and a_i signal coincidences (modified Hebbian learning). Each coincidence increases the respective synaptic strength and each missed coincidence decreases it. When the so-called learning fixation threshold is achieved, the synaptic weight turns to one and the associative connection between the main signal *s* and the associative signal a_i is established. Thereafter a number of these associatively connected a_i signals may evoke the output *so* alone (associative evocation) and they may amplify the output if the main signal *s* is present (priming function).

Associative evocation as described here can only switch on the *so*-signal. Sometimes switching off or inhibition is needed. This can be effected via the threshold control. The output signal duration is normally limited by a decay process. This decay is followed by a short refractory period during which no output is possible. Decay is not applied to direct sensory signals.

Distributed main signal arrays are processed with blocks of these neurons in parallel, usually with common associative inputs. "Winner-Takes-All" (WTA) thresholds at the neuron outputs may be used to select the strongest signals. Sequential circuits may be assembled by using additional short-term memories, delay-line type or other. Strategies to eliminate associative interference e.g. the exclusive-or problem exist (Haikonen 1999b).

4. Perception/Response Reentrant Loop

A cognitive system uses perception processes to access information about its environment and its own physical states via sensors. The perception of an entity in the cognitive sense does not primarily involve the recognition of a pattern; it involves the evocation of the purpose, significance, name, etc. potentially everything that is associated to the sensed signal arrays. The interpretation of these sensed signals to represent one object and not another, to have one set of associations win over others, perhaps equally or even more probable from the sensory point of view, depends on the experience and contextual state of the cognitive system. Thus the whole cognitive capacity of the system is available to assist the perceptive recognition process.

The Perception/Response Reentrant Loop is devised by the author as the basic system module that performs the functions of sensory perception, establishment of inner representations; inner imagery, inner speech etc., introspection, reverberating short-term working memory and the generation of response.

The perception/response reentrant loop consists of a feedback neuron group with output threshold, association neuron blocks and a related Winner-Takes-All neuron group. The signal array at the output of the feedback neuron group is labelled as the percept. It is the official output of the loop and is broadcast as such to other loops and to the pleasure/displeasure system. The percept may be the preprocessed input signal array as such, input signal array primed by the feedback signal array, feedback signal array as such or a combination of the sensory input signal array and feedback signal array.

The feedback neuron group consists of one neuron for each input line. Each neuron has one associative input that receives its signal from the respective association neuron block output WTA neuron so that the inherent meaning of the main signal and the associative feedback signal are the same.

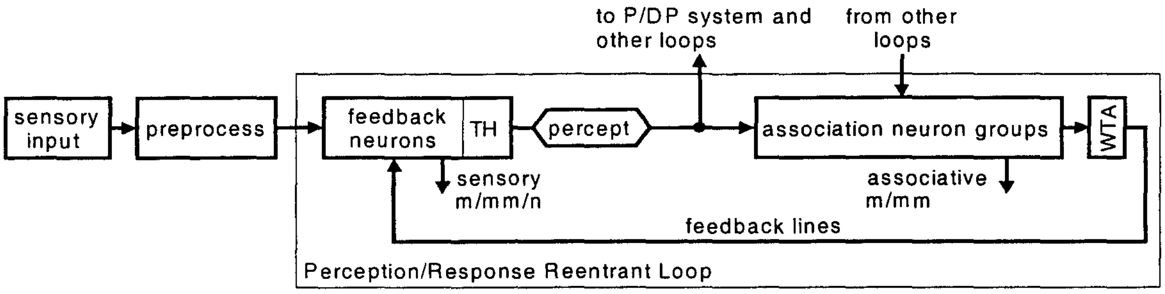


Fig. 4.1. The Perception/Response Reentrant loop

Perception process combines the effect of sensed distributed signal representations and internally generated representations. This can be achieved by inserting neurons into the attribute signal lines and using the associative inputs as reentry points for feedback signals from the system. Initially when the system is in unlearned state, there is no feedback and the sensed distributed signal representation passes through the feedback neurons as such and will be the percept for the system. However, when the system is learned, feedback signals may be generated. In this case *perception with priming* takes place. The percept will now be a non-linear sum of the sensed signals and the feedback. Thresholds may now be applied at various points within the system so that only the thus amplified part of the sensed signal array will have effect.

Two cases of perception with priming can be distinguished. *Predictive perception* occurs when the feedback signals represent a prediction or expectation of the input. This prediction may arise due to associations to previous percepts within the system. What is expected will be more likely perceived. Match/mismatch states derived from the feedback neurons indicate the accuracy of the prediction. *Searching perception* occurs when the feedback signals are an internally evoked representation of a desired entity to be found or distinguished. In that case the match/mismatch states indicate the successful/ unsuccessful status of the search.

The reentry mechanism facilitates also introspection. *Introspective perception* of inner imagery or other inner representations takes place when the percept is due to the feedback signals only, when there is no sensory input or the input is subdued.

The use of the terms *inner imagery*, *inner speech* etc. can be justified as follows. In introspective perception the feedback signals are translated into the percept signals. These signals have the same point-of-origin meaning as the sensory attribute signals, which in turn are in causal connection to the sensed external world. In visual domain these signals represent visual attributes or features and the percept signals represent a sensed image of the external entity. Therefore, whenever any percept signals are evoked internally, it is as if the respective visual attributes were sensed; from the system's point of view the situation is equivalent to that when an image is sensed. However, in this case there is no sensed external entity, the image is internally evoked. These inner images may not necessarily contain all the attributes of sensed external world, they may not be as vivid. These images are not necessarily those sensed before, instead novel images are possible due to the nature of distributed representation. Therefore the term "inner imagery" may be used here. The same line of arguments applies to other sensory modalities as well.

Eventually the system must be able to distinguish between true sensed imagery and internally generated inner imagery, otherwise it would react to internally generated imagery as if it were of external origin. The following cues are available: The activity status of sensory preprocess, *mm*-signals, intensity, vividness.

5. The Cognitive System

A modular system architecture that is based on the previously presented principles is described here. This system architecture consists of perception/response reentrant loop modules for linguistic, visual and visual attention focus position (gaze direction) subsystems as well as subsystem modules for pleasure/displeasure and match/mismatch evaluation. The modules are associatively crossconnected so that the percepts from individual modules can be globally broadcast allowing cross-associative operations.

The linguistic system perceives input words, associates percepts from other units to words, associates words to words, evokes inner representations for words and perceives them as words via the reentrant feedback and enables the flow of inner speech. Words are represented as distributed letter signals so that each individual letter is represented by one on/off signal. As words are, in principle, temporal sequences of phonemes or letters in this case, the letter signal representations are transformed into parallel form so that the letters of the represented word always set in fixed positions and are available simultaneously.

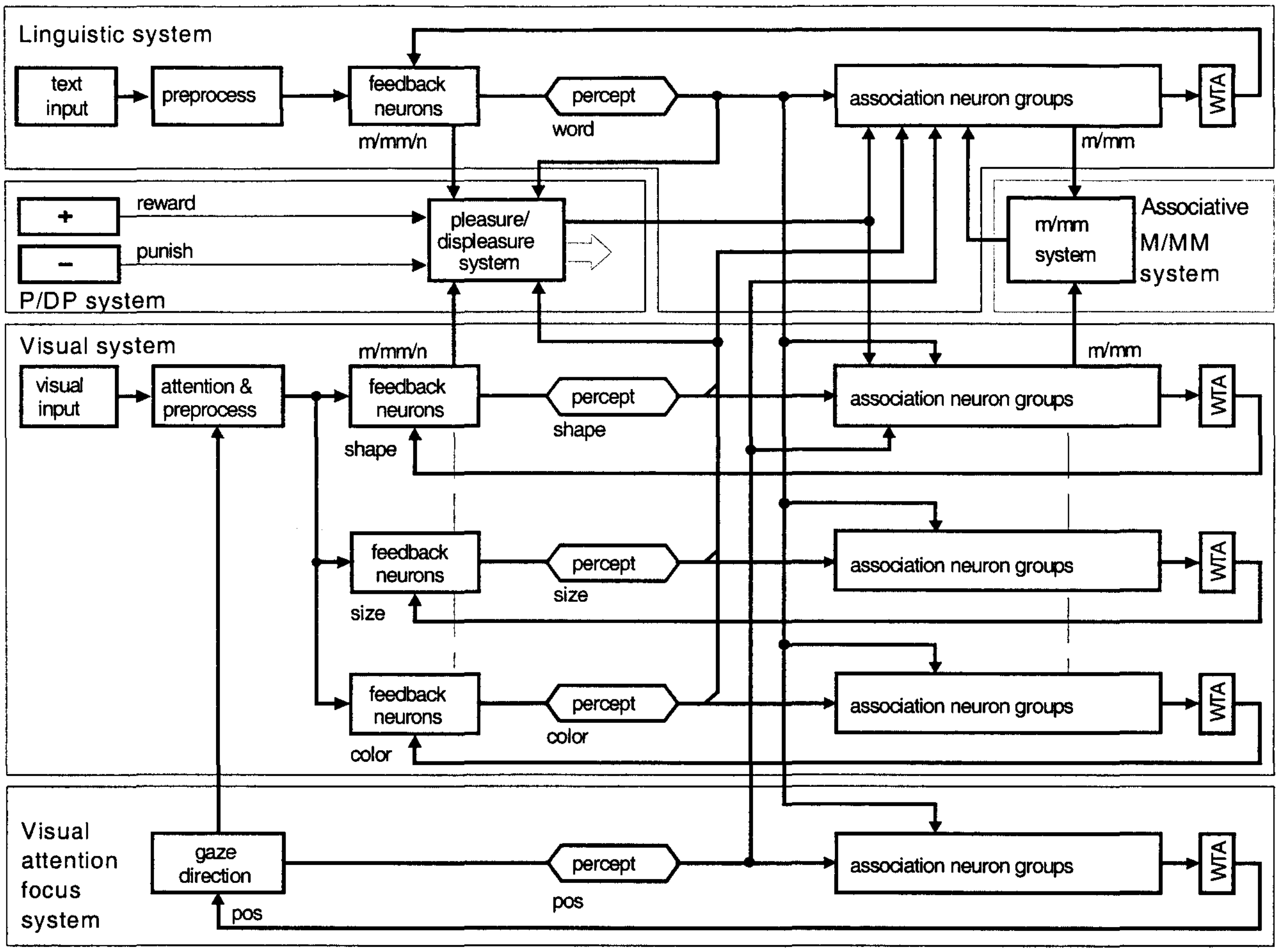


Fig. 5.1. The Artificial Cognitive System

The visual system perceives sensed visual objects; patterns with color and size, associates percepts from other modules to visual objects, associates visual objects to other visual objects, evokes representations of visual objects internally by percepts from other modules and perceives them via associative feedback and in this way enables the flow of inner imagery. The visual system supports the two-way meaning process for concrete words, perceived words with visual meaning evoke the inner representation of the respective visual object.

The visual system preprocesses images into distributed feature signals for shape, color and size. Each of these features has its own perception/response reentrant loop. Images are not reconstructed at any point, the system does not internally operate with actual images. The binding of the features of each recognized entity takes place automatically via associative amplification and thresholding.

The visual attention focus position system controls the focus position and temporarily associates visual objects to their positions. Visual attention focus position is determined by visual change and internally by associative evocation. Visual change is detected by the visual preprocessor.

The associative neuron groups generate match/mismatch signals that indicate the match/mismatch relationship between percepts and evoked representations. These signals are used for learning control and the indication of affirmative and contradictory states, which is important for deduction and reasoning.

The pleasure/displeasure system (P/DP system) associates good/bad significance to percepts and guides judgement, motivation and attention. There are two input sources for the P/DP system. 1.) External reward and punishment signals are accepted, 2.) Match-states from feedback neurons generate pleasure and mismatch-states generate displeasure.

Reward- and punishment-related pleasure and displeasure can be associated to percepts so that later on similar percepts will evoke respective pleasure/ displeasure signals. These percepts are now said to have p/dp significance (functionally more or less similar to emotional significance in biological systems). This significance will translate into signal intensity whenever the said percepts arise again. This signal intensity in turn will guide attention via WTA-circuits and other thresholds so that the significant signal arrays will gain priority (focussing of attention and priming of perception due to emotional significance.)

Linguistic input and visual percepts may be associated together by repeated coincidences. Thereafter sensed visual features will evoke the best matching word, which in turn will evoke the respective visual feature signals at the visual association neuron groups. These will be then prime the visual perception process via the reentry loop and those sensed visual feature signals that belong to the evoked visual entity will be amplified. The same amplifying process can be due to the p/dp significance and other possible sensory modalities. In this way the perception process is primed by the system's knowledge and instantaneous state. In the same way words will evoke respective visual feature signals even if nothing is visually sensed. These evoked inner representations will be sustained by the visual reentrant loops for a while thus enabling introspection so that e.g. questions about them may be answered by the linguistic system. In this way via the cross-associative links the loop modules have access to the other reentrant loops and are thus able to report the contents of the other loops in their own terms –words, imagery, etc. and are also able to affect the contents of the other loops.

6. Simulation System

The artificial cognitive system has been simulated with a system consisting of a computer program and a digital camera.

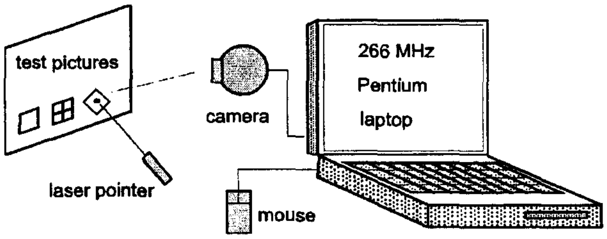


Fig. 6.1. The simulation system

The simulation system hardware consists of a 266 MHz Pentium II laptop computer with mouse, digital color camera and laser pointer to point test figures. The computer operating system is Microsoft NT 4.0. The software consists of a C-language interface for the camera control and a Visual Basic language program for the modular neural system simulation and the user interface. These programs communicate via dynamic link library (DLL) functions.

The camera view is processed into a total visual area and a much smaller visual attention focus area which can move within the total area. Illumination change is detected within the total visual area and actual imagery is detected within the visual attention focus area. The position of the visual attention focus area is determined either by an illumination change (illumination by the laser pointer) or by an internal command from the neural network.

In order to limit the number of required neurons the total visual area is limited to 66 x 66 pixels and the visual attention focus area to 16 x 16 pixels.

When a figure is detected the preprocessing program fine-tunes the visual attention focus area position so that the figure is centered automatically.

The visual object perception and recognition is based on detected features. As a rather limited recognition power is enough for the demonstration of cognitive principles, the number of derived signals is reduced in order to limit the number of required neurons. However, the derived features must enable size invariant recognition, for instance a square must be identified as a square regardless its size. Also minor distortions in the figures must be tolerated.

For the purpose of feature detection the visual attention focus area is divided into four quadrants and a center cross area. Four features; one horizontal, one vertical and two diagonal short lines can be detected within each quadrant. In addition two horizontal and two vertical lines can be detected within the center cross area. Each feature is represented by one signal, visual figures are represented as arrays of feature signals.

Words and sentences may be entered from the keyboard at any time. Each word may contain up to six letters. Each individual letter is represented by one on/off signal. No words are stored at any point as alphanumeric strings.

The simulation program starts in unlearned condition, therefore usually some figures, sizes, colors and their names are taught first by ostension and then higher concepts are taught by example sentences. Thereafter simple conversation is possible. The simulation program learns fast, e.g. the meaning for a word can be taught in few seconds. When the desired effect has been completed the cognitive processor window and the camera window may be captured by the print screen command for documentation.

The following cognitive functions and processes have been demonstrated among others:

Perception with priming; predictive perception. This involves the evocation of a continuation to the perceived state of affairs, a rudimentary deduction process. The generation of prediction match/mismatch value and corresponding match/mismatch pleasure/ displeasure is also included.

Perception with priming; searching perception. This involves the priming of perception with the representation of the searched item and the generation of match signal when the item has been found.

Introspective perception. This involves the perception of the system's inner responses.

Visual sensory attention. The ability to focus the gaze on individual visual objects. In this simulation gaze direction is controlled by external stimulus, i.e. visual change caused for instance by a laser pointer, and by internal command. Additionally visual sensory attention is related to the perception of individual visual feature signals. The attended signals will be the strongest ones, due to external or internal causes.

Inner attention. Inner attention is based on the operation of the various WTA neuron groups that pass the strongest signals. Priming, decay etc. affect the strength of individual signals.

Learning concrete words by ostension. Ostension involves the pointing out the intended item and associating a word to it. This is also a two-way process, afterwards the item must be able to evoke the given word and the given word must be able to evoke the inner representation of the respective item.

Learning concrete words via correlation. Sometimes the item to be named cannot be pointed out exclusively. In that case different items with the desired attribute as common can be used as examples.

Learning by examples. Learning category names by example question-answer pairs ("what color red", "what name candy" etc.) and learning the meaning of "yes" and "no" and to recognize a question to which yes or no is expected as an answer, again by example question-answer pairs ("is this square yes", is this candy no") has been demonstrated. There is also some inductive power, the learned examples can be used in different context.

Learning of rudimentary syntax. The ability to recognize a question and to answer to it properly involves the learning of rudimentary syntax.

Learning by verbal description. Learning by verbal description is one form of social learning and it involves the availability of common language between the learner and the teacher. In this simulation system inner representations for new visual objects can be created by verbal description using already known visual objects as components and a name can be associated to these objects ("small green square dollar"). Thereafter whenever the new object is actually imaged it will be recognized ("what name *dollar*").

Learning of sequences. The learning of sequences involves temporal association. This is realized here by sequential predictor circuitry within the association neuron groups and its output is perceived via the reentrant loop.

Sensory Match/mismatch/novelty signals. These signals indicate whether there is match or mismatch between the sensed representation and the internally evoked representation or whether the sensed representation is novel.

Affirmation and contradiction detection. This function is based on associative match/mismatch detection, which matches representations against each other whether evoked externally or internally. This is a prerequisite for reasoning by inner imagery and is also needed for the grounding of meaning for words like "yes" and "no".

Pleasure/displeasure function and p/dp significance. Pleasure and displeasure are system's reactions that try to sustain the prevailing attention and activity or disrupt and refocus them. In this simulation system pleasure/displeasure signals are used to initiate actual reactions. Due to the limited scope of the simulation system the actual reactions are rather limited. E.g. when visually searched object has been found, match pleasure is generated and this in turn will hold the visual attention on that object. *The p/dp-significance* manifests itself in elevated signal levels, which e.g. facilitate the detection of p/dp-significant patterns.

Short-term memory via loop reverberation. Short-term memory is needed for working memory. The perception/response reentrant loops are able to sustain representations via reverberation.

Long-term memory via synaptic weights. Long-term memory is based on the accumulation and fixation of synaptic weights.

7. Conclusions

Do we have here a mind, a mind with its own joys, sorrows, secret desires, perhaps even with existentialist suffering? Definitely not. But we do have an architecture, mechanism and platform that supports some of the prerequisites of the mind; the operation with meaning and significance, unification of information from multiple sensory modalities and internal knowledge, learning, the flow of inner imagery, inner speech, introspection, sensory and inner attention. Cognitive processes involving the utilization of inner imagery and inner speech have been demonstrated.

The demonstration of anything like goal-oriented and emotionally motivated action would necessitate the inclusion of motor output (even if virtual), needs, and provision for system reactions. It remains for future research to see if and how emotional states could be emulated with total system reactions and attention control induced by the P/DP-system.

The author has presented here one possible way how a cognitive neural system can be actually assembled. The simulated system is at this moment admittedly limited in scope but it could be easily expanded. A more complete cognitive system for actual applications (e.g. robotics) would include further sensory modalities (tactile etc.) and motor outputs. Also more extensive serial/parallel capacity should be included in the association neuron groups.

The author's system can be compared to other models of mind and consciousness. Baars' global workspace theory (Baars 1997) proposes a "theater stage" as the site for inner imagery and inner speech. In the author's model the percept locations may be compared to the Baars' "theater stage" as they contain the inner speech and inner imagery and these representations are broadcast to the other parts of the system. Baars proposes that this "theater stage" is located at the sensory projection areas, which is also the case in the author's model and which is also quite necessary for the grounding of the basic meaning for the inner representations. However, Baars does not really explain how information should be represented by neural firings or how actual neurons should be connected into networks that would constitute a complete cognitive system.

How about machine consciousness? Self-consciousness is not yet emulated here, as the simulation system does not have episodic memory for personal history nor body reference for self-concept ("I") and therefore is not able to perceive itself as the executing agent. Even though the system has the flow of inner speech and inner imagery and it operates with them, it is not yet able to report having them. It is not able to produce much towards the response "I have inner imagery" or the consequence "I think – therefore I exist". Obviously this kind of a report would only count as a proof of self-consciousness if it can be seen that the system is producing it meaningfully, i.e. the system would have to be able to perceive its inner imagery as such and it would have to posses the concepts like "I", "to have" and "inner imagery". The mere reproduction of preprogrammed strings like "I have inner imagery" would not count as a proof here. -The author would like to see the Turing test (Turing 1950) be replaced by this one. May the race begin!

References

Aleksander Igor (1996), *Impossible Minds My Neurons My Consciousness*, Imperial College Press, London

Ashcraft Mark H. (1998), *Fundamentals of Cognition*, Addison Wesley Longman Inc, New York

Baars Bernard J. (1997), *In the Theater of Consciousness*, Oxford University Press Oxford New York

Devlin Keith (1997), *Goodbye Descartes*, John Wiley & Sons, Inc. New York

Haikonen Pentti O. A. (1998a), "Machine Cognition via Associative Neural Networks" in *Proceedings of Engineering Applications of Neural Networks EANN'98* Gibraltar 10 - 12 June 1998 pp. 350 - 357

Haikonen Pentti O. A. (1998b), "An Associative Neural Model for a Cognitive System" in *Proceedings of International ICSC/IFAC Symposium on Neural Computation NC'98* Vienna 23 - 25 September 1998 pp. 983 - 988

Haikonen Pentti O. A. (1999a), Finnish patent no 103304

Haikonen Pentti O. A. (1999b), *An Artificial Cognitive Neural System Based on a Novel Neuron Structure and a Reentrant Modular Architecture with Implications to Machine Consciousness*, Dissertation for the degree of Doctor of Technology, Helsinki University of Technology Applied Electronics Laboratory, Series B: Research Reports B4, 1999

Hinton Geoffrey E., McClelland James L., Rumelhart David E. (1990), "Distributed Representations" in *The Philosophy of Artificial Intelligence*, pp. 248 - 280, Margaret A. Boden, editor, Oxford University Press, New York

Lenat Douglas B. (1995), "Artificial Intelligence", in *Scientific American*, September 1995, pp. 62 - 64

Nairne James S. (1997), *The Adaptive Mind*, Brooks/Cole Publishing Company USA

Omar Rosli (1994), "Artificial Intelligence through Logic?" in *AI Communications* Vol 7 Nos 3/4 pp. 161 - 174

Searle John R. (1984), *Minds, Brains & Science*, Penguin Books Ltd, London England

Taylor J. G. (1999), *The Race for Consciousness*, A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England

Trehub Arnold (1991), *The Cognitive Brain*, MIT Press, London England

Turing Alan M. (1950), "Computing Machinery and Intelligence" in *Mind* LIX, no 2236 Oct. 1950, pp. 433-460

Valiant Leslie G. (1994), *Circuits of the Mind*, Oxford University Press, Inc, U.S.A

How to make a Self

Pat Hayes
Institute for Human and Machine Cognition
University of West Florida http://www.coginst.uwf.edu/~phayes

Abstract

The computational paradigm can account, in broad terms, for many of the phenomena of consciousness, as Dennett (1991) argues convincingly. The general idea is that the mechanism of consciousness is an internal representational narrative (also called a 'global workspace' or a 'world model'), suitably embedded in a functional architecture, and the phenomenal aspects of consciousness are the content of this narrative. However, this kind of account has some problems, most notably the fact that there seem to be aspects of this internal narrative which we are not, and cannot be, conscious of. The distinction between conscious awareness and unconscious information processing (for example in the visual system) must therefore be based on something more than the simple presence of the relevant information. Another missing aspect is the self; the sense of personal integrity which is a hallmark of normal conscious experience.

Building on pioneering work by Perlis (1997), we will suggest a representational solution to both of these problems. Perlis posits an "ur-quale" which arises from a particular kind of self-modelling computation. This paper sketches how strong self-reference might evolve in a simpler architecture as a result of the interaction of processes of truth maintenance, episodic memory and active spatial location, and why the resulting representational structure would give rise to many of the characteristic phenomenal aspects of the subjective self. This account has the merit of not requiring exotic computational architectures to support a self-concept. Finally we will suggest ways in which this mechanism might break down and produce psychoses.

References

D.C. Dennett, *Consciousness Explained*, Back Bay, 1991

D. Perlis, Consciousness as Self-function, in *J. Consciousness Studies* 4, 5-6, pp 509–25, 1997

A design study for an Attention Filter Penetration architecture

Brian Logan
School of Computer Science & IT,
University of Nottingham, Nottingham UK.
bsl@cs.nott.ac.uk

Abstract

This paper describes a design study for a variant of the 'Attention Filter Penetration' (AFP) three layer architecture (Sloman, 2000). In an AFP architecture, the activities of an agent are distributed across three concurrently executing layers: a *reactive* layer in which detection of internal or external conditions immediately generates new internal or external response, a *deliberative* layer responsible for 'what if' reasoning and planning capabilities, and a *meta-management* layer which provides self monitoring, self evaluation and self-redirection including control of attention. Our design is based on two main assumptions: that deliberation (and meta-deliberation or management) is the technique of last resort for an agent, and is only used when no reactive behaviour is clearly applicable in a given situation; and deliberation is just something that some reactive systems do. The paper attempts to identify those parts of the design which seem fairly uncontroversial, and to highlight those issues which have yet to be resolved.

This paper describes a design study for a variant of the 'Attention Filter Penetration' (AFP) three layer architecture (Sloman, 2000). The aim is to design an architecture for an agent which can control its efforts to achieve one or more goals in some domain, and for the system to be able to adapt its behaviour to changes in the environment and difficulties encountered in achieving its goals.

In an AFP architecture, the activities of the agent are distributed across three concurrently executing layers: a *reactive* layer in which detection of internal or external conditions immediately generates new internal or external response, a *deliberative* layer responsible for 'what if' reasoning and planning capabilities, and a *meta-management* layer which provides self monitoring, self evaluation and self-redirection including control of attention. An attention filter with a dynamically varying interrupt threshold protects the resource-limited deliberative and meta-management layers when dealing with tasks that are important, urgent and resource consuming.

We make two main assumptions: that deliberation (and meta-deliberation or management) is the technique of last resort for an agent, and is only used when no reactive behaviour is clearly applicable in a given situation; and deliberation (and management) is just something that some reactive systems do. In doing so, we are not attempting an explanatory reduction the deliberative or management layers to the reactive layer, rather the aim is to sketch an implementation of the virtual machines which operate at these layers in terms of the primitives available at the reactive layer. Some such reduction must be possible: some kinds of deliberative and management behaviour must 'just happen' otherwise we end up with infinite regress. However there is no reason in principle why they should reduce to mechanisms at the reactive layer, they could, for example, be implemented using distinct machinery 'at' their respective layers. The assumption that they are not is perhaps the central design decision of this paper.

Of particular interest therefore is the interaction between the reactive and deliberative layers: both the generation of new motives or goals by the reactive layer and when and how these are scheduled for processing at the deliberative layer. We focus first on the reactive layer of the architecture to clarify what it can and can't do, and to outline how it does what it does, before attempting to show how the deliberative and management layers can be implemented as particular kinds of reactive behaviour.

The paper attempts to identify those parts of the design which seem fairly uncontroversial, and to highlight those issues which have yet to be resolved. In several cases, a number of possible approaches to an unresolved issue are identified. Such speculations are not intended as exhaustive enumerations of the options, rather they attempt to indicate the current state of work on a topic and illustrate insofar as this is possible at this stage, some of the main issues that would have to be addressed by any solution.

1 The Attention Filter Penetration architecture

In this section, we briefly describe the Attention Filter Penetration three layer architecture which forms the basis of this study.

The Attention Filter Penetration architecture attempts to account for the existence of a variety of more or less sophisticated forms of information processing and control in human and other minds. The version discussed here is based on previous work by Sloman and others (Sloman and Croucher, 1981; Beaudoin, 1994; Sloman, 1994; Sloman and Poli, 1996; Sloman, 1997, 1998, 2000; Sloman and Logan, 1999) and postulates three concurrently active layers which evolved at different times and are found in different biological species. The three layers account for different sorts of processes. None of the three layers has total control: they are all concurrently active and can influence one another.

The first layer contains *reactive* mechanisms which automatically take action as soon as appropriate conditions are satisfied. The second *deliberative* layer provides 'what if' reasoning capabilities, required for planning, predicting and explaining. The *meta-management* layer provides the ability to monitor, evaluate, and partly control, internal processes and strategies.

Roughly, within the reactive layer, when conditions are satisfied actions are performed immediately: they may be external or internal actions. A reactive system may include both analog components, in which states vary continuously, and digital components, e.g., implementing condition-action rules, or various kinds of neural nets, often with a high degree of parallelism.

By contrast, the deliberative layer, instead of always acting immediately in response to conditions, can contemplate possible actions and sequences of possible actions (plans), compare them, evaluate them and select among them. The human deliberative systems can also consider hypothetical past or future situations not reachable by chains of actions from the current situation, and can reason about their implications. Deliberation requires a large amount of stored knowledge, including explicit knowledge about which actions are possible and relevant in various circumstances, and what the effects of various actions are in those circumstances. It also requires a re-usable short term memory for building structures representing possible action sequences in order to evaluate their consequences. The reuse of memory, the inability of the long term store to answer many questions in parallel, and the sequential nature of plan construction will typically make a deliberative system much slower than a reactive one.

The meta-management layer acts on some of the internal processes involved in the reactive or deliberative (or meta-management) system. This includes monitoring, evaluating and redirecting such internal processes, and possibly reflecting on them after the event in order to analyse what went wrong or how success was achieved. Like the deliberative layer, it will be resource-limited.

Both in a sophisticated reactive system and in a deliberative system with planning capabilities there is often a need for motives which represent a state or goal to be achieved or avoided. In simple organisms there may be a fixed set of drives which merely change their level of activation depending on the current state of the system. In more sophisticated systems not all motives are permanently present, so there is a need for *motive generators* to create new goals possibly by instantiating some general goal category (*eat something*) with a particular case (*eat that foal*). These generators are similar to the dispositional 'concerns' in Frijda's theory (Frijda, 1986). Beaudoin (Beaudoin, 1994) and Wright (Wright, 1997) discuss various types of generators or 'generactivators' and related implementation issues.

Since the different layers operate concurrently, it is possible for new information that requires attention to reach a deliberative or meta-management sub-system while it is busy on some task. Because of resource limits, the deliberative sub-system may be unable to evaluate the new information while continuing with the current task. However it would be unsafe to ignore all new information until the current task is complete. New information therefore needs to be able to interrupt deliberative processing.

Under stressful conditions, deliberative mechanisms with limited processing or temporary storage capacity can become overloaded by frequent interrupts. Beaudoin & Sloman (Beaudoin and Sloman, 1993) have argued that a variable-threshold attention filter can reduce this problem. Setting the threshold at a high level when the current task is urgent, important and intricate, can produce a global state of 'concentration' on that task. Conversely, malfunctioning of this mechanism may produce a type of attention disorder (Beaudoin, 1994).

2 The reactive layer

In this section we present some observations about the nature of reactive behaviours and sketch a design for the reactive layer of a simple agent.

We assume that the agent has some simple behaviours which are triggered by the presence or absence of certain features in the environment. Some of these behaviours are atomic, whereas others are composed of simpler behaviours and may have complex internal organisation such as conditionals, loops etc. which *implicitly* anticipate the future. In general behaviours have duration; some behaviours are effectively instantaneous, but for those that are not, it is often more natural to talk in terms of an action such as *move forward* rather than a sequence of equivalent actions, such as *move forward 1m*, each of which can be accomplished within some time bound. In many cases, multiple behaviours can run in parallel. Some behaviours which can't simply be executed in parallel, for example because they require the use of one's hands, can however be 'blended'; in other cases the activation of two behaviours for the same processor or which require the same 'resource' may give rise to a explicit goal to choose between them (see below).

Usually such behaviours are completely automatic; the presence of their triggering condition(s) in the environment will always elicit the behaviour. However, the execution of these behaviours is subject to various forms of control. Some 'reflex' behaviours can be temporarily suppressed or their execution modified as a result of feedback from the environment. In some cases, complete suppression of the behaviour requires 'prior notice', e.g. not blinking while undergoing an eye examination, though some modification is often possible even after the behaviour has been initiated, e.g. stifling a cry of alarm or surprise. Other forms of control appear to be global modifiers of all behaviours at the reactive layer, for example when late for an appointment we may perform all routine actions hurriedly, or aggressively when we are frustrated. Such control also extends to modifying the expression of a voluntarily executed behaviour or set of behaviours, as when we perform an action 'carefully', 'quietly', 'theatrically' etc. In this case, it is not clear if the control is accomplished by 'stepping through' the basic actions of the behaviour in such a way as to change their subsequent execution, or whether we actually synthesise a new 'careful' version of the behaviour 'on the fly' from more basic behaviours.

It is convenient to model reactive behaviours as sets of condition-action rules. The condition part is matched against the agent's perception of the current world state, and, if it fires, it triggers a single (ballistic) action which attempts to change the state of the world in some way. Such a behaviour is entirely stateless, in that it maintains no internal representation of the state of the world, whether the rule has been fired before etc. However, even such basic 'reflex' behaviours can be chained together to produce quite complex behaviour, for example, if we (or evolution) arrange things so that one action changes the world in such a way as to trigger another action, and such an architecture can support conditionals and simple loops.

One problem with this approach is that if any action fails to achieve its 'intended' result, then the whole sequence of actions may fail unless: (a) the action had no significant effect, allowing the rule to be re-invoked in the hope that this time it will succeed; or (b) the action has some other effect, recognisable as an 'error condition' and there is an 'error recovery rule' that can get things back on track for the 'intended state'.

Another problem is that the agent will always respond to the same situation in the same way, and will continue to respond in the same way if its efforts to change the world are unsuccessful. One way round this is to arrange for the rules to match against some internal representation of the environment, rather than percepts directly generated by the agent's sensors. This allows the agent (or some of the agent's behaviours) to respond only to *changes* in the environment, ignoring those features that are constant (without some representation of the previous state, we can't say what is novel in the current state). One way to build such a representation is to use simple percept driven condition-action rules to record simple 'beliefs' about the state of the world. Such rules perform a very primitive kind of belief fixation, in which the 'beliefs' are simple flags in the agent's internal state which correspond one to one (modulo sensory noise and other errors) with the world as the agent perceives it.[1] Other rules match against this internal representation, and, if they fire, trigger actions. From the point of view of expressive power, this adds nothing. We have simply taken a condition-action rule and split it in two, with a mediating internal representation. To detect and represent changes in the environment, we also need rules whose conditions match against the agent's internal state and whose actions modify the agent's internal state.[2] Given such internal behaviours, much more complex external behaviours are possible.

More complex reactive behaviours also require additional internal representations to hold intermediate state during their execution, for example unachieved implicit goals or the current state of a task (where 'state' is understood as a predefined stage in a task, rather than some general representation of the world). This includes negative feedback loops (where there is at least an implicit representation of the state to be achieved or maintained). As with the simple reflexive behaviours described above, this kind of architecture supports conditionals, looping etc., though in this case we can loop a fixed number of times without having to record anything in the external environment, but the sets of rules can be more naturally described as simple control programs rather than collections of simple reflexes.

Although reactive behaviour is essentially data-driven, it may be useful to conceive of it as being goal-directed. We distinguish between two types of goals: implicit and explicit. Implicit goals are not subject to deliberation. We can view a reactive behaviour which is triggered by, e.g., an approaching object (and hence the possibility of a collision) as a response to an implicit goal of avoiding collisions. For purely data-driven reactive behaviours, this way of looking at things adds nothing, but it provides a uniform interface to behaviours which can either be triggered by events in the world or under the control of the agent. In particular, it gives us a way of indexing reactive behaviours by their effects, so that we can find behaviours appropriate to a deliberative goal.

An *explicit* goal is generated whenever the reactive layer doesn't know what to do. We assume that the agent has some sort of control program or reactive behaviour for all classes of events which are of interest to the agent. Percepts or representations of the world are matched against the agent's reactive behaviours, and some or all of the behaviours which match are invoked. Any 'percepts' or

[1] Specifically it is not intended to be a general representational capability, and cannot handle beliefs about things which are not present to the agents senses (unless they were previously), or indeed about many of the things which are. However it could be viewed as the first step towards such a representational capability.

[2] An arrangement in which 'world-to-representation' and 'representation-to-novelty' rules is replaced by a single rule with conditions which match both percepts and internal representations would of course also work, but the notion of a rule which only responds to and generates internal changes in the agent is a key step, since it forms the basis of all derived representations, and of representations which refer to other aspects of the agent's internal state.

sense data which don't trigger some sort of behaviour are simply ignored by the agent. This is perhaps more credible if we assume an agent with several levels of perceptual processing, where a failure to produce any sort of perceptual classification for a for low-level percept, e.g. hearing an unfamiliar noise, may give rise to an explicit goal, e.g. to investigate the source of the noise. If a percept matches a behaviour but the behaviour cannot be executed (perhaps because a precondition is not satisfied), or the behaviour fails during execution and the reactive layer is not able to recover, then an explicit goal is generated by some error handling mechanism in the behaviour. Similarly, if a percept matches several behaviours, this could also give rise to a deliberative goal. We may want to assume some sort of heuristic partial order over behaviours, or many such orderings (e.g. effectiveness, reliability, speed etc.), with different orderings being used in different situations. Only in the case in which there is no clear preferred behaviour is an explicit goal generated. The result is an architecture which is basically reactive, in which deliberation is used as a last resort. The agent only engages in explicit deliberation about what to do next when a reactive behaviour 'fails' in some way. What happens when a new explicit goal is generated is discussed in more detail in section 5.

3 The deliberative layer

An explicit goal is one which involves (explicit) deliberation, for example the explicit generation and evaluation of alternative courses of action by the agent, or a decision to perform some action. Such explicit goals can be either *intrinsic* or *derived.* An intrinsic goal is one which is not a subgoal of an already intended end. A derived goals is a subgoal generated in response to an existing intrinsic or derived goal.[3] Some steps in the deliberation may take the form of basic actions (possibly expressed as implicit goals, see above), which simply trigger some reactive behaviour; for example, adding two single digit integers, or comparing two alternative actions to decide which is preferable.

Beaudoin and Sloman (Beaudoin and Sloman, 1993) identify ten fields associated with an (explicit) goal including: a possible state of affairs p; a propositional attitude towards p; a value representing the agent's current beliefs about p; an importance value relating to the benefits of achieving the goal or the costs of not achieving the goal; the urgency of the goal; a plan or set of plans for achieving the goal; a commitment status such as 'adopted', 'rejected' or 'undecided'; and management or scheduling information in the form of condition action pairs, determining, for example, when execution of the plan or further deliberation should begin or be resumed if it is currently suspended.[4] The urgency and importance of a goal can be thought of as modifiers of the basic propositional attitude ranging along two (possibly orthogonal) dimensions. The degree of urgency or importance of a goal is relative to that of the other goals the agent is currently trying to achieve and therefore can't be determined *a priori.* Goals also have an insistence value which is a heuristic estimate of the goal's likely urgency and importance (and possibly other things), which can be defined operationally as the ability of the goal to penetrate the attention filter. In what follows we will focus on the commitment status, management information and insistence of a goal.

The basic unit at the deliberative level is the *task.* A task is a declarative representation of a sequence of actions to achieve a goal. Deliberative tasks organise or synchronise reactive behaviours in pursuit of an explicit goal, and may involve complex internal organisation and the explicit anticipation of possible futures. In contrast to reactive behaviours they are goal rather than data-driven, and can involve commitment to future action.

A task has several components:

- the goal the task is trying to achieve, including any constraints on how the goal can be achieved;
- when the task has been scheduled, i.e. when we intend to start work on the task; and
- any preconditions that must be true before we can perform the task.

A task can be thought of as stack consisting of an initial intrinsic goal and the unachieved derived subgoals generated in attempting to achieve the original goal. The top of the stack is the current subgoal and the preconditions are the conjunction of the preconditions for the pending actions. Tasks are executed by a deliberative 'interpreter' which is itself simply a collection of appropriately organised reactive behaviours responsible for pattern matching, manipulation of the goal stack, updating working memory etc.

The scheduling conditions of a task define a partial order over tasks which tells the agent which tasks it should be working on at any one time. As used here, 'scheduling' is to be understood loosely as any absolute or relative temporal ordering over tasks, e.g., "I'll do it on Friday", or "I'll do it after I have finished debugging this function". In this model, Bratmans's (Bratman, 1987) notion of a 'committed intention' is equivalent to a task that has been scheduled. A task will not usually be considered for execution until all its scheduling conditions evaluate to true. The rationale for this is that if we have decided that we won't do something until next week, there is no point in continually checking the preconditions for the task to see if it could be executed next. Presumably the task was

[3] Georgeff and Lansky (Georgeff and Lansky, 1986) call these *operational* goals.

[4] Note that the use of some terms, e.g., 'adoption', differs from that in (Beaudoin and Sloman, 1993).

scheduled next week for a reason: for example, it may be that we can't do it sooner, or that we anticipate that we will have more than enough to do in the meantime.

When the scheduling conditions for a task evaluate to true, the task can be considered for execution. A task is executable if its preconditions evaluate to true. By default, new tasks are created with no scheduling information or preconditions (unless the task is a subtask of an existing task, see above). New preconditions are added whenever the agent must wait for some condition to be true in the world before the task can continue, for example for an action to complete (such as travelling to a particular location) or for a resource to become available. The preconditions can be any state in the world or the agent and will usually be dependent on the level of the task. For example, a repair task may be suspended awaiting the delivery of a replacement part, or for a supplier to confirm availability or price of a part, whereas a scheduling (management) task, might be suspended awaiting an estimate of how long the task will take to execute, or whether a particular way of achieving a goal is feasible. If one or more of the preconditions of a task evaluate to false, the task is suspended until the precondition becomes true.

Processing at the deliberative and reactive layers proceeds concurrently. While, the current deliberative task is the focus of the agent's attention, deliberation associated with the current task proceeds in parallel with the execution of actions associated with any suspended tasks by the reactive layer. For example, we may be walking down the corridor in pursuit of an explicit deliberative goal (to get some coffee) while thinking about something else.

4 Managing deliberation

Periodically, the agent must consider whether it should continue with its current deliberative task or switch to another. This is the process loosely referred to as 'scheduling' in the previous section.

The scheduler is an independent process which runs in parallel with the deliberative interpreter, and which continually monitors its progress. When a new goal penetrates the filter, or if the scheduler detects that the current task is not making progress, or the scheduling condition for a higher priority task has just become true, the scheduler arranges for the deliberative interpreter to stop executing the current task and start executing the new task. If making such a decision itself requires deliberation, the scheduler creates a new deliberative scheduling task, and causes the deliberative interpreter to switch to this task.[5]

Scheduling can be non pre-emptive if we can assume an upper bound on the time necessary to execute a basic action, or that the initiation of a durative action causes the task to be suspended while waiting for the action to complete (e.g., (Georgeff and Lansky, 1986)). Alternatively, if actions can take a substantial amount of time to complete or we cannot suspend the current task while the action is executing (for example, if it is not easy to determine a success condition for the action to use as a precondition), then it may be more appropriate to use pre-emptive scheduling (which may result in any work for partially complete subgoals/basic actions being lost).

One factor which is important in determining what to do next is how much progress has been made towards the current goal. A plan may be of the form "do X" (e.g., try and jiggle a part into a socket, or try and find a plan) or of the form "do X for a while, and if that doesn't work, try Y". In the former case, we have to rely on general monitoring by the scheduler to discover that we are not making progress. If a behaviour has an 'outcome' related to the implicit or explicit goal the behaviour is trying to achieve (rather than simply a 'success' or 'failure' flag), then the scheduler can look at the outcome and decide whether to terminate the behaviour, allow it to continue, or suspend it to allow another task to run. One way the current outcome of a behaviour can be determined is to compute the (relevant) change in the world since the behaviour started execution, or to compare the current state of the world with the goal. Information about the current outcome could be used by an explicit monitor for the plan, or could form some sort of annotation on the task to tell the scheduler how long to persist with X before switching to Y (all other things being equal). Like other behaviours, progress monitoring is partially reactive and partially deliberative. Routine monitoring is handled reactively, with goals being generated if the reactive progress monitoring behaviours can't determine if progress is being made.

When the scheduler decides that X is not achieving its goal, we must either replan (if this contingency was not anticipated) or switch to Y (if it was). For example, we could arrange for the scheduler to send some sort of 'failure' control signal to the current task, causing it to backtrack and select another behaviour. In the worst case, this would lead to the failure of the top-level goal for the task, which could trigger replanning. This has the advantage of localising plan repair within the task or plan. However it implies a task representation with an explicit notion of 'failure'. Alternatively, we could exercise control directly at the meta-level. A plan is not a behaviour; it is a declarative structure, the major steps in which are explicitly executed. We could use this representation to reason explicitly at the meta-level (perhaps using task-specific knowledge) about where to backtrack to and which other behaviours might achieve the goal.

We assume that most of the decisions about which task to switch to next can be be made quickly by the data-driven or reactive parts of the scheduler responding to automatically generated meta-level descriptions of the current state of the system's deliberative tasks, and we only use deliberation to determine what to do next as a last re-

[5] The alternative approach of giving the the scheduler its own deliberative interpreter doesn't seem to have the right phenomenology, since we want such deliberative scheduling tasks to interrupt the current deliberative task, whatever it is.

sort.[6] The default scheduling policy may be to stick with the current task unless the scheduling conditions or preconditions of another task have just become true, or we are not making progress with what we are doing, for example if the task is taking longer than estimated, when it may be necessary to generate an explicit management goal. Initially, any unusual situation may give rise to a management goal, but with experience, the agent may develop reactive behaviours for the more common cases, perhaps as a result of a process similar to chunking in SOAR (Laird et al., 1987).

5 New intrinsic goals

As described in the preceding sections, any events which can't be handled by the reactive behaviours of the agent give rise to new intrinsic goals. In this section, we discuss filtering of new intrinsic goals and the process of goal adoption, and argue that the decision whether to adopt a goal can most usefully be viewed as part of the scheduling process.

When a goal is generated by the reactive layer, it must first pass the attention filter to be eligible for consideration by the mechanisms at the deliberative layer. The purpose of the attention filter is to avoid interrupting the current task unless it is likely that the new goal is both urgent and important in the current context. In general, it is impossible to determine the urgency or importance of a goal, and hence if we should consider the goal further, without expending (possibly considerable) deliberative resources. By assumption, such deliberative resources are not available to the filter mechanism, and passing the problem to the deliberative layer would defeat the point of having the filter in the first place. Instead, the attention filter uses simple heuristics based on the goal's insistence and the current filter threshold to determine if the goal should receive further consideration. The attention filter will therefore be prone to 'errors', passing goals which on further consideration are not worth adopting, and failing to pass goals which it would have been beneficial for the agent to have adopted.

Whether an agent should *in fact* attempt to achieve a goal may depend on many factors, and may require considerable deliberation, for example deciding whether one should agree to organise a workshop at a conference or write a paper about a particular piece of work. Consideration of how, when and ultimately whether to intend some state X may be spread over a protracted period, and it is difficult to justify the choice of a particular point as *the* point at which a goal to achieve X is adopted. Rather there is a wide range of intermediate states each of which may affect the way the agent responds to the putative goal and its other tasks in different ways. For example, we may not have decided how urgent or important a goal is, or consideration of the goal may be suspended pending information necessary to make a decision whether to pursue it further (e.g., the information about the cost of travelling to X may be instrumental in deciding whether we will go there), or the goal may have been provisionally adopted but only more or less loosely scheduled, e.g., "I'll do it if I have time", to "I'll do it next week (sometime)", to "I'll do it on Wednesday at 4pm" to "I'll do it next" to "I'll do it now". Moreover, while Bratman (1987) is correct in that we don't continually reconsider our commitment to our existing intentions and that such commitments provide an essential framework for planning, intermediate stages prior to 'full' commitment can still be useful for scheduling. For example, I may decide not to accept a dinner invitation next week because I believe that I have a lot of things tentatively scheduled for next week, and if pressed to make a decision at short notice may conclude that I am unlikely to be able to fit everything in. Given more time to decide, I may have been able to schedule all my current intentions and still accept the invitation. In what follows, we assume that new goals are automatically 'adopted', at least provisionally, after passing the attention filter, and that a new task is created for them.

Once it has been adopted, the processing of a new intrinsic goal is similar to the scheduling of existing tasks. The first time the scheduler runs after the adoption of a new intrinsic goal, there is a new task with no information associated with it, for example, the task goal will typically have no urgency or importance fields. This lack of information makes it difficult to use heuristics to decide whether to switch to the new task (and more generally what to do with it, e.g., whether it should be deleted), unless either the new task or the current task is extremely urgent and/or important, causing all other tasks to be ignored. For example, the sound of a fire alarm and the resulting goal to leave the building immediately may be sufficiently unambiguous that it takes precedence over all other deliberative tasks. Usually however, the failure of the heuristic scheduling behaviours result in the generation of a new explicit management goal to find out more about the task to determine if we should switch to it. This may involve working out how the task can be done, how urgent and important it is (relative to the other tasks the agent currently intends) and so on, and ultimately results in a heuristic or deliberative scheduling decision.

At first sight, it may seem that new management goals must be handled in a slightly different way. If we try to decide what to do next and can't, we generate a goal to think about it. This goal is like a new intrinsic goal in that it is generated reactively; moreover it is not a derived goal of any of the existing tasks. However, it seems unreasonable that a management goal should have to pass the attention filter, as if it *failed* to pass the filter (with the result that we fail to attend to the goal) we still wouldn't know what to do next. Similarly, if it passes the filter but isn't adopted or we don't switch to processing it, we are

[6] If several scheduling rules apply, we could pick one at random or allow them to 'vote' for which task to run next. If they disagree, then this might be another reason for generating a meta-management goal.

still stuck. Management goals must therefore have special status within the architecture.

However, the context in which a management goal is generated is usually such as to ensure that it will be processed next anyway, without having to take special measures. If it is not clear what to do next, the filter threshold is presumably not very high, since the presence of an urgent and/or important task (which could easily be scheduled reactively) implies a high filter threshold. If management goals generally have a high insistence value, there is at least a good chance that a management goal will be selected as the current deliberative task at the next iteration of the scheduler. This approach has the advantage of allowing preemption of management goals by new, urgent 'ordinary' goals—the context changes and it suddenly becomes clear what we should do or think about next—and it may make it easier to explain certain types of perturbances where meta-management fails.

6 Summary

In this paper, we have sketched a design for the reactive-deliberative interface in an Attention Filter Penetration architecture. In doing so, we have tried to explain how new explicit goals are processed within the architecture and how it is that some goals are capable of redirecting the attention of the deliberative system.

Working within the framework outlined in (Sloman, 2000), we have sketched the progress of a new explicit goal from its initial generation in the reactive layer in response to a reactive failure, through the attention filter and into the deliberative layer, and briefly described how the new goal interacts with existing goals at the deliberative layer. We have focussed on the role of the scheduling mechanism which decides which goal to work on next, and argued that scheduling, like the other behaviours of the agent, is partly reactive and partly deliberative. Reactive scheduling failures give rise to management goals. At first sight, it may seem that such goals must have a special place within the architecture. However, we have argued that the context in which a management goal is generated is usually such as to ensure that it will be processed next anyway, without having to take special measures. This allows preemption of management goals by new, urgent 'ordinary' goals—the context changes and it suddenly becomes clear what we should do or think about next—and it may make it easier to explain certain types of perturbances where meta-management fails.

We have left many important questions unanswered. Further work is required to clarify (among other things) the nature of the deliberative-reactive interface: how the deliberative layer can invoke reactive behaviours and modulate their execution. Humans seem to be able to execute and modify a wide range of reactive behaviours under deliberative control, descending recursively through the component behaviours of a compound behaviour to obtain the required degree of control. For example, we may execute some steps in a plan essentially unconsciously, while closely controlling other steps, or parts of steps. More work is also required to elaborate the detailed behaviour of the attention filter and scheduler. This is the subject of current research.

Acknowledgements

Aaron Sloman, Luc Beaudoin and Ian Wright read an earlier version of this paper and made many helpful comments.

References

Luc P. Beaudoin. *Goal processing in autonomous agents*. PhD thesis, School of Computer Science, The University of Birmingham, 1994.

Luc P. Beaudoin and Aaron Sloman. A study of motive processing and attention. In A. Sloman, D. Hogg, G. Humphreys, D. Partridge, and A. Ramsay, editors, *Prospects for Artificial Intelligence*, pages 229–238. IOS Press, Amsterdam, 1993.

Michael E. Bratman. *Intention, Plans, and Practical Reason*. Harvard University Press, 1987.

Nico H. Frijda. *The emotions*. Cambridge University Press, Cambridge, 1986.

M. P. Georgeff and A. L. Lansky. A system for reasoning in dynamic domains: Fault diagnosis on the space shuttle. Technical Report Technical Note 375, Artificial Intelligence Center, SRI International, Menlo Park, California, 1986.

J. E. Laird, A. Newell, and P. S. Rosenbloom. SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33:1–64, 1987.

A. Sloman. Explorations in design space. In A.G. Cohn, editor, *Proceedings 11th European Conference on AI, Amsterdam, August 1994*, pages 578–582, Chichester, 1994. John Wiley.

A. Sloman. What sort of control system is able to have a personality. In Robert Trappl and Paolo Petta, editors, *Creating Personalities for Synthetic Actors: Towards Autonomous Personality Agents*, pages 166–208. Springer (Lecture Notes in AI), Berlin, 1997.

A. Sloman. Damasio, Descartes, alarms and meta-management. In *Proceedings International Conference on Systems, Man, and Cybernetics (SMC98)*, pages 2652–7. IEEE, 1998.

A. Sloman and M. Croucher. Why robots will have emotions. In *Proceedings of the 7th International Joint Conference on AI*, pages 197–202, Vancouver, 1981.

A. Sloman and R. 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.

Aaron Sloman. Architectural requirements for human-like agents both natural and artificial. (what sorts of machines can love?). In Kerstin Dautenhahn, editor, *Human Cognition And Social Agent Technology*, Advances in Consciousness Research, pages 163–195. John Benjamins, Amsterdam, 2000.

Aaron Sloman and Brian Logan. Building cognitively rich agents using the Sim_agent toolkit. *Communications of the Association of Computing Machinery*, 42 (3):71–77, March 1999.

Ian P. Wright. *Emotional agents*. PhD thesis, School of Computer Science, The University of Birmingham, 1997.

Shakespeare's invention of theatre as simulation that runs on minds

Keith Oatley
Centre for Applied Cognitive Science
Ontario Institute for Studies in Education-University of Toronto

Abstract

In any fully functioning cognitive system, such as the human mind, emotions would be central since, as Simon (1967) pointed out, either they, or something like them, are needed to manage cognition and action. The management is in relation to an outer world of objects and events for which mental models will always be incomplete and sometimes incorrect, and for which agency will often be inadequate. In the human case it is also in relation to an inner world in which we humans have many goals (concerns), some of which are in conflict with each other, and in relation to a social world in which we cooperate and conflict with other agents constituted somewhat as we are.

I propose that the central issue of designing a complete cognitive system relates to this last issue: distributed cognition and agency. We humans bridge our cognitive deficit of inadequate knowledge and agency, by cooperating with others to extend our mental models and capabilities. We are members of that species who accomplish together what we cannot do alone. This is the solution to which the evolution of the human brain has devoted most of its computing resources.

A principal means of improving our understanding of such matters is indeed simulation, but—perhaps paradoxically in the context of AISB—the kind of simulation that runs on minds rather than on computers. In modern Western culture, it was Shakespeare who first implemented this idea. Shakespeare's great innovation was of theatre as a model of the world. The audience member constructs the simulated model in the course of the play, and thereby takes part in the design activity. So fiction is to understanding social interaction as computer simulation is to understanding, perception and reasoning. Shakespeare designed plays as simulations of human actions in relation to predicaments, so that the deep structure of selfhood and of the interaction of people who have distinct personalities becomes clearer. I explore this idea by analyses of *Henry IV Part 1, As You Like It,* and *Hamlet.* As we run such simulations on our minds, we not only construct and experience the emotions of the vicissitudes that cause them, but we are enabled to reflect on them to create deeper mental models of individuals (including ourselves) and of interaction. Understanding the properties of such mental models is the principal step in designing a fully human-like mind.

Simulation and the role of emotions in distributed cognition

If we should want to simulate a working cognitive system in something like the quotidian world that we human beings inhabit, emotions would be central. Emotions can be thought of in as processes that manage cognition and action in the individual mind. The management is, as Simon (1967) pointed out, in relation to an outer world of things and events for which our mental models are always incomplete and often incorrect, and for which our agency is frequently inadequate. It is in relation to an inner world in which we have many goals and concerns some of which are incompatible with each other. It is also (as Simon did not point out) in relation to a social world in which we cooperate and conflict with other agents who are constituted somewhat as we are (Oatley, 1992).

Emotions in the individual are types of readiness for certain repertoires of action, and they are experienced as urges towards these actions. Aubé and Sentenyi (1996) have called them commitment operators. Not only do they make the cognitive system ready to act in a certain way, they commit us to act in this way. When angry, for instance, an individual becomes committed to redress and finds is hard to think of anything else.

More important than individual psychology is the consideration of how emotions work in social interaction. Emotions are commitments in the social world—the world of distributed cognition and action. Anger properly then is the emotion of social conflict, of one person getting his or her way in relation to another person who does not.

Evolution, emotions, sociality

I propose that the central issue of designing a complete cognitive system relates to distributed cognition and agency. Outline scripts (role-relationships) for social interaction are based on emotions. Jenkins & Greenbaum (1999), and Oatley and Jenkins (in press) have proposed three primary kinds of socio-emotional scripts, each with its prototypical emotion and its higher level goal.

The first is attachment of the kind described by Bowlby (1971). It emerged in evolutionary time some 70 million year ago, and it is characteristic of mammalian life. It probably forms the foundations of cooperative sociality. Its prototypical emotion is anxiety and its overall goal is protection (Goldberg, Grusec, & Jenkins, 1999) in the first place of infants from predation and intra-species aggression. Later in life it provides bases for trust and protective attitudes towards others more generally, or alternatively for its opposite: distrust and Machiavellianism.

The second is assertion. In evolutionary time this may have emerged earlier than attachment, in relation to dominance hierarchies which are widespread in mammalian and avian social species. Its characteristic emotion is anger. In evolution its overall goal has been to challenge for, or to respond to a challenge, for position (status) in a dominance hierarchy. This is the system for within-group conflict and competition. In humans the concern with status is typically felt in terms of self-esteem. Anger sets up a frame, a script, or as Averill (1982) says a temporary role, for a status dispute. It occurs typically in a relationship with someone close (a son, a daughter, a spouse, a colleague) when something such as a slight or a failure to do what was promised undermines mutual expectations in the relationship. If one is angry, the emotion commits one to seeing the issue through socially, renegotiating status—who was to blame, who should apologize, who should undertake some amendment of life—and typically also coming to some reconciliation, which usually involves readjusting the relationship in order to continue on somewhat different terms.

The third is affection. This system is distributed much more unevenly among mammals. It is widespread only in primates, and it becomes distinctive only in humans. Its emotion is affectionate happiness, and its overall goal is cooperation. In humans this system became especially important with long-term sexual relationships in which the male made an economic contribution to the rearing of specific offspring, beginning some 3 to 5 million years ago (Lovejoy, 1981).

We are members of that species who cooperate to accomplish what we cannot do alone. The system of cooperation is the means by which we humans bridge the cognitive deficit of our inadequate knowledge and agency. By cooperation with others we extend our mental models and capabilities. Moreover we can sometimes deal with multiple goals by having different people represent different concerns.

Conversation and the origins of language

We infer from the work of Dunbar (1996) that cooperation, in a background of competition, is the solution to which the human brain has devoted most of its computing power. Primates have relatively larger brains than other mammals. The human brain is some 1300 cc in volume; that of a typical non-primate mammal of our body weight is 180 cc. The increase in size largely accounted for by the neocortex, which reaches 80% of brain volume in humans. The larger cortical size in primates is correlated with fruit eating and foraging over large territories. Most of all, however, it is associated with living in highly interactive social groups.

Aiello and Dunbar (1993) have shown that the ratio of neocortical size to the rest of the brain in primates correlates closely with the mean size of the social group of that species. The two species with the largest brains are chimpanzees and ourselves. Here is Dunbar's core hypothesis. The increase in size of primate brains, as one moves from species to species, is based on the number of others for whom one not only has individual mental models, but models of pair-wise relationships in the group. Chimpanzee social networks can include some 60 individuals, known in what we might call this personal way. Human social networks have about 140 or so, perhaps upto 200.

Dunbar has also proposed that cooperation in primate groups is maintained by mutual grooming: the basis of affectionate friendships and alliances. Many species of monkeys and apes spend up to 20% of their time in grooming. But as group size increases, so does the number of affectionate relationships one needs to maintain, and hence the amount of grooming one needs to do. With a mean group size that is more than twice that of chimpanzees, humans would need to spend more than 40% of their time grooming. Add another 33% for sleep, and one can see that time left for the business of acquiring food and other necessaries of life shrinks to an implausibly low amount.

Dunbar's solution is that that, in humans, conversation has taken over the function of grooming. On the basis of estimated group size and the amount of grooming required to maintain one's affectionate relationships, Aiello and Dunbar calculate that human language emerged about 250,000 years ago. Conversation enables us to not only to do something else while we verbally groom, but also to groom with more than one person at a time. (Manual primate grooming is with only one other individual at a time, and it excludes other activities.) Perhaps most importantly—though Dunbar does not discuss this—perhaps in the explicit commitments that are made to other people, to joint plans, to shared beliefs, conversation is probably more efficient in forming and maintaining relationships than manual grooming. In studies of what people talk about, in conversations Dunbar (1996) has found that, indeed, some 70% of it is about the social lives of ourselves and our acquaintances. These people are friends and enemies, the trusted and the untrustworthy. What is known as gossip, the informal recitation and analysis of action and character, is about forming mental models of others, and of pairwise relations between others, with whom we have interacted in the past, and with whom we may interact in the future

Here, then, is my hypothesis, following Dunbar. Suppose that in evolution the hypertrophy of the human cortex was due to developing mental models of a large number of others, and that conversational language evolved to maintain cooperative relationships. If, moreover, as argued above, the real stuff of life is emotional rather than perceptual interpretation and intellectual problem solving, then if we wish to model a fully human-like cognitive system, we should follow the lead of primate evolution. We should devote the larger part of computational resources to social issues.

One could argue that even if Dunbar is right, and that hypertrophied human brain is based on sociality and conversation, that once modules for "mental-model-of-the-individual" and "mental-model-of-the-dyadic-relationship" and perhaps some higher order group models were properly designed, they would merely need to be iterated to deal with all the people we know. No new principles need be evolved.

This may be partially correct, but it is not likely to be fully correct for three reasons. First is the large difference in brain size as a function of body weight between non-primate mammals and primates, for instance between members of the cat and dog families and monkeys. Cats and dogs are clearly social, but in ways but that do not include mental models of distinct individuals, as monkeys and apes do. Second is the fact that language is a complex function to which substantial neural computing power is devoted. Third is that, both in everyday human conversation and in the problems that continue to preoccupy us, it is the social world that fascinates. It remains intensely important to us: whom should we chose as a sexual partner, how can we continue when someone of immense importance to us has died, how can we respond to a child who seems uninterested by school and chooses the most unsuitable friends, how should we treat a colleague who is so cantankerous as to make dealing with him/her impossible?

Artificial Emotion (AE)

Among the considerations we need to design a fully functioning mind are, if I may use a notion coined by E.O. Wilson (1998) not just artificial intelligence (AI) but its counterpart artificial emotion (AE). The emotional themes I discussed above—attachment-based anxiety, assertion-based anger, and affection-based happiness—become the bases for principal themes. It is the possibilities of enactment of these themes among individuals who have different characters—bold, boastful, manipulative, affectionate, envious, obsessive—that has posed during evolution, and continues to pose in our everyday lives, the great challenge to our brain-power and our mental models. It is this that leads to the indefinitely large number of scenarios and plot lines that we try to understand.

We should, moreover, remember, as Neisser (1963) pointed out, that in humans adult mental life is based on accretion within evolutionary and developmental sequences. The propensity to trust or distrust, for instance, is thought to be based on species-specific attachment schemas, and on early experience in at least one attachment relationship. Assertion-based aggression is based on temperament and the particularities of success and failure in status disputes. Amiability and cooperativeness are though to be based on temperamental warmth and on a person's history of affectionate relationships.

My proposal is that within human culture, and within the developing minds of individuals, a principal means of improving our understanding of such matters is by means of simulations. But, perhaps paradoxically in the context of an AISB conference, these simulations run on minds rather than on computers. I propose that in modern Western culture, Shakespeare was the first to grasp this idea fully.

The simulations of fiction

Fiction no doubt arose from conversation—from the "She did such and such, and then do you know what happened?" This is the basis. It grew to group story-telling, and oral recitations.

The devices of oral recitation in small groups were augmented by two developments that extended to larger groups: religious rituals and the invention of writing. In the two Middle-Eastern/Western cultures that have preserved a continuity of written language for 3000 years—Hebrew and Greek—we see how these streams have developed in somewhat different ways. In Judaism, the written word became the central element in worship. The togetherness of worship involved (and still involves) both rituals and cultural commitment to ideas and ideals. In Greece, the oral tradition of storytelling, rendered into written form by Homer, also had a quality of ritual meetings in substantial numbers. It came to be implemented in the theatre, where the narrator was replaced by an actor in counterpoint with a chorus. The theatre grew to exceed the temple, in size and perhaps in importance.

Shakespeare's great innovation, soon after the even more expansivist invention of printing, was his idea of theatre as a model of the world. The audience member constructs a simulated model in the course of the play, and thereby becomes part of the design process. One sees Shakespeare's idea, and its aspiration to human universality, not just in his calling his theatre "The Globe." One sees it not just in speeches like "All the world's a stage," in which a metaphor (the literary term for simulation) transforms our vision of an aspect of reality. One sees it in the deep structure of his plays. So theatre and fiction are to understanding social interaction as computer simulation is to understanding perception and problem-solving.

Shakespeare's plays are simulations of the interactions of people with their predicaments so that the deep structure of selfhood and social interaction becomes clearer. Shakespeare's idea was to take seriously what Aristotle called *mimesis.* I have argued (Oatley, 1992, 1999) that *mimesis* is best translated as simulation. The simulations that are plays and novels run on minds rather than computers. Many of the considerations of computational simulations apply also to literary ones. For instance, in computation there are two kinds of code. Some code represents aspects of the real world that is being simulated. Other parts are instructions to the computer about how to conduct the simulation. Similarly in any story or play, one aspect which we may call the story structure (that Russian literary theorists called the *fabula*) is representation of the story world. The second aspect, the discourse structure (*siuzhet*) has attributes of speech acts, instructions to the reader or audience as to how to construct and run the simulation. As we run such simulations on our minds, we not only experience the emotions and hence the urgency of the human vicissitudes and dilemmas that cause them, but we are enabled to reflect on them in such a way as to create deeper level mental models of ourselves and others.

As with any simulation, literary simulation selects some aspects as important, and these are emphasized by being the ones that are set into interaction with each other, to produce the outcomes. What, then, are the main aspects of Shakespeare's simulations? I argue that there are three.

Three aspects of Shakepearian simulation

The first is the basic structure of all narrative: goal-directed actions by human agents who meet vicissitudes. Here begins the emergence of mind, in our compulsion to see action as purposeful. Narrative is the computational language for expressing such purposes, and within it the structure of causation in both the interpersonal and physical world. "She was so angry that she took all his stuff out into the garden and made a bonfire of it." The reader/listener infers, and mentally constructs, the causal sequence: purpose—>action—>outcome-with-the-aid-of-the-physics-of-combustion.

The second is often argued to be Shakespeare's invention: his depiction of what in literary theory is called character, and in psychology is called personality; see, for instance, Bloom (1999). Here is the idea in cognitive terms. People's actions and thoughts flow from interpersonal goals that are habitual, and hence somewhat predictable by self and others. We define character, and its effects, in terms of such habitual goals. As Henry James (1884) put it: "What is character but the determination of incident? What is incident but the illustration of character?"

The third has been widely recognized but not so explicitly discussed. It is the aspect of emotions. We all know that fiction includes the idea that emotion occurs when a human goal meets with a vicissitude. I propose, in addition, that further sense can be made by means of the three primary interpersonal motivational systems—attachment, assertion, affection—which are inherently emotional. Character then becomes, in part at least, a predominance of one of these emotion-based systems that has, in the better kind of fiction, another in conflict with it. So, for instance, Hotspur in the Shakespeare play that I discuss first, is impulsively aggressive. He is the Renaissance warrior full of derring do. But he is also an affectionate husband. The effects of incident upon the individual (character, audience member, reader) then, become the somewhat habitual emotional responses of the individual to vicissitudes (incidents) in interpersonal contexts.

Three plays

Let me now discuss three of Shakespeare's plays in the order in which they were written, between 1598 and 1600, *Henry IV Part 1, As You Like It,* and *Hamlet,* to show his development of his idea of mimesis-simulation, and some of his implementations using these three aspects.

Henry IV Part 1

The first of these plays, *Henry IV Part 1,* (Shakespeare, 1623c) is about politics. Politics is about assertion, about getting one's own way, but it is also about the other two socio-emotional modes: attachment-based trust and distrust, and affection-based cooperation.

Surface and deep structure. Those of us who have had occasion to observe or to take part in politics, either at the national level or in smaller contexts such as the university, will recognize the following paradox. Politics is that domain in which, because we have limited mental models but nevertheless need to act, is typically guided by cooperative group discussion (Aristotelian dialectic), with one person or faction arguing for doing this and another for doing that. It occurs in the context of status-based hierarchies. On the surface, in political discussion, people give reasons in the rhetorical form of rationality. But one may think, on listening to them, that their arguments are tendentious: motivated by desires that are kept beneath the surface. There is an equivalent surface level in oneself: "I am right and they are wrong." In such discussions one may then hear oneself being authorititative, shrill, or hectoring. The paradox is that as one begins to take part in the discussion, the same doubt as to tendentiousness, as to good versus bad faith, is immediately cast by others upon oneself. They may then cease to listen to anything one has to say. Political discussions, thereby easily become not so much explorations about how best to act, but competitions based on the assertive social modes of power, solidarity, and fear.

So, how should the essayist, playwright, or novelist portray the deep structure of politics? To denounce is to adopt the same genre of discourse, to take up a role indistinguishable from politicians on whom one wishes to comment. The answer comes, I suggest via Erasmus, the central literary figure in the northern Renaissance, who wrote about 500 years ago.

The influence of Erasmus. Without Erasmus there would have been no Shakespeare. Erasmus was the first writer to benefit largely from the invention of printing. It was he who set the curriculum for school-based education as reading and writing: hence the grammar school in Stratford that Shakespeare is thought to have attended, in which he would have practiced the classical-medieval mode of putting an argument, and then with equal force its antithesis (see, for instance, how he handles the debate about slaying Caesar, in *Julius Caesar*).

Erasmus's books were widely read. We know, from his use of them, that Shakespeare read Erasmus's collections of *Adages* (Latin translations of Greek sayings and quotations). An example is: *"mare malorum,"* "sea of troubles," which occurs in the most famous speech in all of Shakespeare: "To be or not to be ... to take arms against a sea of troubles ..." (*Hamlet*, 3, 1, 56-59).

It must be certain that Erasmus's most popular book, *Praise of Folly*, (1508) would have been read by the intensely bookish Shakespeare. In *Praise of Folly*, Erasmus introduces a metaphorical figure Folly—a woman in that age of male public action—and has her give a speech in praise of herself: a foolish thing to do. In this book Erasmus proposed that what is on the surface is not typically what is important. His way of doing this was to personify the disowned (seemingly foolish) emotional aspect of public life, together with instructions (*siuzjet*—discourse structure) to the reader to run the simulation in the mode of irony. The effect is to prompt one toward forming a representation that includes both emotions and reasons, with the irony prompting reflection not only on the deep structure of many kinds of public discourse, their pomps and their circumstance, but on one's own involvement in such structures. By taking part in such simulations, one might even come to prefer the candidness of emotion over the more dignified discourses of so-called reason.

My suggestion is that it was *Praise of Folly* that prompted Shakespeare's crystallization of his idea about how to portray the deep structure of politics as a stage-based simulation of several characters in interaction. *Henry IV Part* was his first implementation.

High life simulated by the low. In the opening speech of the play King Henry proclaims that now civil strife in his kingdom has ended, he and his men at arms can go, no longer divided but united, to undertake a Crusade, to make war on foreigners. The speech is stirring and patriotic: or (not an exclusive "or") under the ironic lens that Shakespeare has inherited from Erasmus, it is an utterance that betrays the King's compulsive purpose of aggression.

Then comes news of more civil conflict, and the (regretful?) postponement of the foreign adventure to put down an incursion in Wales. Quickly thereafter comes news of the young Hotspur who has routed a Scottish force and taken many prisoners, valuable for purposes of ransom, a source of aristocratic booty. Like the King, Hotspur represents assertion-aggression, and the King voices his regret that his own son and heir, Prince Hal, is not more like him. The King also represents distrust (attachment based) and the calculating manipulativeness to which it gives rise—a character trait that will emerge also in his son who later becomes Henry V, in the play of that name.

By means of alternating scenes, Shakespeare juxtaposes (simulates) the aristocratic group by means of another group: that of the dissipated Jack Falstaff and his layabout tavern friends who include Prince Hal. In a deeper and more psychological simulation, Shakespeare presents Falstaff as surrogate father to the Prince.

In the second scene, Falstaff and his gang also hatch a plan of force of arms: a highway robbery of some pilgrims. Poins and Prince Hal subsequently conspire to arrive late for their arranged part in the robbery, and then to rob Falstaff and his three companions of the booty, so as to witness Fastaff's cowardice and subsequently to hear his lies as he recounts the episode. Here are the seeds of the young Prince's manipulativeness.

Jacobson (1988) has proposed two basic modes of language: juxtaposition, the metonymic, and seeing a as b, the metaphorical. Shakespeare is the master of both. Here he uses the one to accomplish the other. This play is the first in theatrical history to juxtapose depictions of high and low life on the stage. By this means Shakespeare achieves metaphor in the large: aristocratic politics as gang-based brigandry.

These devices, which extend throughout the play, also include the opportunity for Erasmusian irony: Falstaff, after he too has taken a cowardly part in the Battle of Shrewsbury—the Falstaff whom many audience members have come by this juncture to like—speaks ironically of the machinery of political force, honour:

> Can honour set a leg? No. Or an arm? No. Or take away the grief of a wound? No. Honour hath no skill in surgery, then? No. What is honour? A word. What is in the word honour? What is that honour? Air. A trim reckoning. Who hath it? He that died o' Wednesday (5, 1, 131-136).

The device of simulating the ennobled by means of the dissipated also enables the transformation of Prince Hal from the role of tavern layabout to his proper role as His Majesty, in a way that retains his character intact. At the end of Act 1, scene 2, he offers a soliloquy about his intended metamorphosis that is both self justifying and coldly calculating. By the end of this play he becomes the princely Prince who fights and kills Hotspur in the Battle of Shrewsbury. Two plays later he is the kingly King Henry V.

Elizabethan politics. It is said that it is hard to gauge Shakespeare's political sympathies. He is thought to have been born into a Catholic family (though evidence is ambiguous because it needed to be kept quiet for the political reasons of the very kind that we are discussing). He is known to have moved to London which was largely Protestant at that time, not long after the Catholic-Protestant antagonisms that still bedevil us had been born—a time also when that other figure who influenced him much, Christopher Marlowe, was killed for political reasons.

I believe we understand Shakespeare's political stance better if we see him as intensely fascinated by, but also horrified by, public violence. He strove to depict not whom to take sides with, which is the easy but violence-reproducing option for us all. He strove to depict the very stuff of politics. He represented a world in which Henry V became the great English national hero who triumphed at Agincourt against those traditional enemies, the French, who was at the same time the ruthless and manipulative leader who used self-serving rhetoric to mobilize his troops, and a foreign war to stifle criticism at home. (Thus has British politics continued into recent times.) Such portraits cannot be sustained by the either-or of political discourse. They need simulations of people whose different, and conflicting, parts of character are brought into action as individual centres of consciousness that purposefully affect external events, and bring about the vicissitudes of their own plans. Such a simulation is not an invitation to believe this or do that. The audience or reader must make the final integration by running the simulation on her or his own mind.

As You Like It

In *As You Like It,* (1623a) the play in which the lugubrious Jacques offers his speech "All the world's a stage, I And all the men and women merely players," Shakespeare tips us off as to his method and intent.

Simulation within the play. The idea is clear enough. It is reiterated elsewhere, for instance in the sorcerer Prospero's speech in *The Tempest,* (1623d) which is often seen as the farewell to the London stage by Shakespeare, the ultimate dramatic sorcerer. Here, he not only explains his method, but encloses it in a miniature mimesis-simulation-metaphor of his most universal commonplace, the passing of human life after its brief drama:

> These our actors,
> As I foretold you, were all spirits and
> Are melted into air, into thin air.
> ... the great globe itself,
> Yea, all which it inherit, shall dissolve,
> And like this insubstantial pageant,
> Leave not a wrack [wreck] behind. We are such stuff
> As dreams are made on, and our little life
> Is rounded with a sleep (4, 1, 148-158).

But it is not just this kind of magical invocation that is at issue. The point, as in the *Henry IV* and *Henry V* plays, is the deep structure. In *As You Like It,* the structure at issue is that of cooperative affection and courtship. In this mode, once again the surface is misleading.

Let me put the problem and its paradoxical qualities in prose. Even in the mode of most affectionate cooperation, all is not straightforward because of the enormity of the life-time commitment that is implied by falling in love. To fall in love, the first stage is to be open to the experience and to see someone whom one likes, perhaps accidentally. Then, after an interval when one reflects and builds fantasies about the person, one needs to see the person again. At this time, by means of a word or sign, one hopes for a piece of evidence that one's interest is reciprocated. If it is, then one is in love: there! Yet more importantly, one is reassured that the love is mutual. The difficulty is that, despite the suddenness of such changes intermediate steps must be gradual, and often ambiguous. Such steps need both to be interpretable as signs of intense interest but also, if not reciprocated—such is the delicacy of selfhood—as something quite different. Here is the paradox: the very moment when one must be most ambiguous, hence potentially deceptive, is that in which one must be most open. It would be the worst possible start of this all-important relationship to be dupicitous.

Shakespeare handles this as follows, by means not just of a simulation, which is the play itself, but by emphasizing that what the audience is watching is a simulation. This he does by embedding, within the play, yet another simulation. In *As You Like It,* first he has Rosalind pass, without much ceremony, through the first stage of falling in love with Orlando, and he with her, after Orlando attracts notice by winning a wrestling match. Next, Shakespeare moves the action from the normal world (a ducal court) to an imaginary (simulated) idyllic world of nature, the Forest of Arden. Here, Rosalind dresses up as (simulates) a young man, Ganymede. By this device, when Rosalind (Ganymede) meets with Orlando, he can speak to Ganymede of his love for Rosalind. She is ironic about the state of being in love in general. She promises to cure him of it by acting as young men believe women do act in this state, "proud, fantastical, apish, shallow, inconstant ..." and so forth through many confusing and conflicting moods. To accomplish this she suggests that Orlando should speak to Ganymede as he would to Rosalind: "call me Rosalind, and come every day to my cot [cottage], and woo me." Orlando says he does not want to be cured but, fascinated by the challenge, he does agree. After another interval, he presents himself. He arrives late, so Rosalind reproves him. For lovers, she explains, being even a fraction of a minute late is subject to the severest of interpretations. She says she would "as lief be wooed of a snail."

> *Orlando* Of a snail?
>
> *Rosalind* Ay, of a snail; for though he comes slowly, he carries his house on his head—a better jointure [joint property], I think, than you make a woman. Besides he brings his destiny with him.
>
> *Orlando* What's that?
>
> *Rosalind* Why, horns, which such as you are fain to be beholden to your wives for. But he comes armed in his fortune, and prevents the slander of his wife.

Most of us would be content with the small but conventional joke about a snail's slowness. But one may gain a glimpse of Shakespeare's genius by how he handles it, in his mode of metaphor-simulation. His joke about jointure indicates that Rosalind knows Orlando is a second son. It is also one of which sociobiologists nearly 400 years later would have been proud. Shakespeare goes yet further, elaborating the theme of what men generally think about women using the idea of horns—the badge of cuckoldry—and elaborating too the overall theme of candour in affectionate relationships.

A few lines later Rosalind/Ganymede has pardoned Orlando, and there follows this:

> *Rosalind* Now woo me, woo me, for I am in a holiday humour, and like enough to consent. What would you say to me now an I were your very, very Rosalind?
>
> *Orlando* I would kiss before I spoke.
>
> *Rosalind* Nay, you were better speak first, and when you were gravelled [stuck] for lack of matter you might take occasion to kiss (4, 1, 62-69).

And so forth. It is a delicate scene, full of further erotic wit on Rosalind's part, in which the audience knows, and knows that Rosalind knows, and suspects also that Orlando must suspect, that he is indeed talking to Rosalind, and she to him. All is within a structure of simulation where the indirection allows them both to be direct, as they could never be in the ordinary world beyond the Forest of Arden. From this, might we see our way to being more direct in our dealings with those we love?

Hamlet

Hamlet, (1623b) is a play about grief, the emotional mode that occurs when someone dies, who is an attachment figure held in affection. The question, "To be or not to be," is whether to enter the mode of assertive revenge or to take to suicide.

The play within the play of *Hamlet* is a mimesis-simulation performed by a troupe of travelling actors. Hamlet hatches his purpose to present this simulation of the purposeful killing by Claudius of his father in order to usurp his own succession to the throne and to wed his mother. So he inserts a dozen or sixteen lines into a play which the travelling actors have in their repertoire: "The play's the thing | Wherein I'll catch the conscience of the King [Claudius]." Thereby he plans for his suspicions from the unreliable source of his father's ghost to be confirmed or refuted, and for Claudius's emotional reaction to the simulation to become a public demonstration. Thus is drawn a nice thread between the medieval notion of acting vengefully on mere private suspicion, and the modern notion that one needs publicly accountable evidence to accomplish justice.

In *Hamlet,* however, Shakespeare's mimesis-simulation idea goes yet deeper. For here he returns again to the issue of public violence, with which he dealt in *Henry IV.* But now the scope is widened from Kings and thieves, to everyman, to ourselves when whatever is emotionally closest to us—outrage at the murder of our dearest, our rights being usurped, disgust at our mother's sexuality—takes possession of our deepest and most urgent concerns. In this simulation, each member of the audience becomes Hamlet, becomes depressed, is driven half mad ("but mad north-north-west"), becomes violently aggressive, becomes contemptuous not just of one beloved but of all.

As such we experience the forces driving towards conventional outcomes: suicidal violence against self or vengeful violence against another. Now, in slowed-down paces, Shakespeare steps us through the experience of Hamlet's states (simulated: ours but not ours) so that we the audience experience its emotions and its confusion as the complacent world of family unravels. We become suspicious at being watched, angry at the mother, despairing at being unable to act in tune with our carefully constructed sense of self. None of the ordinary modes in which we are practiced, trust or distrust, assertiveness, affection, is of any avail. We reflect upon the vulnerability of our human existence, even when supplied with the very best of minds as Shakespeare's is, and Hamlet's is.

Conclusion

The issues of social interaction that Shakespeare treats in his simulations are, I believe, those to which the great weight of neo-cortical computing power has been devoted in the species *Homo sapiens* and those to which, if we wished to design a fully functioning cognitive system, we too might properly attend. These issues are of understanding humans such as ourselves, and humans who are different from ourselves, issues of emotions, issues of what goes on between and among people. These are the issues that the human mind is most adapted to compute over.

This adaptation has generated extraordinary abilities to understand, and take part in, both cooperation and conflict, which have been the bases of our success as a species. At the same time, in many of their aspects, the implications of these issues are too difficult for the unaided human mind to comprehend—hence the evolution of the pre-simulations of conversations in distributed cognition, and of the more elaborate simulations of plays and novels, in which the issues can be explored more deeply. Hence, the almost unbelievable fact that by means of a few thousand pieces of language code (words) a whole world, with people in it, can be programmed and summoned up.

By such means we can derive insights: new pieces of code and representation that we can program into ourselves to make sense of human social-interactive complexities. Insights can occur when we both experience emotions relevant to events and simultaneously can interpret these same events. But despite such insights, full understanding seems often to lie just beyond the horizon of human mental modeling.

References

Aiello, L. C., & Dunbar, R. I. M. (1993). Neocortex size, group size, and the evolution of language. *Current Anthropology, 34,* 184-193.

Aubé, M., & Senteni, A. (1996). Emotions as committments operators: A foundation for control structure in multi-agents systems. In W. V. d. Velde & J. W. Perram (Eds.), *Agents breaking away: Proceedings of the 7th European Workshop on MAAMAW, Lecture notes on artificial intelligence, No. 1038, p. 13 –2.* Berlin: Springer.

Averill, J. R. (1982). *Anger and aggression. An essay on emotion.* New York, NY: Springer.

Bloom, H. (1999). *Shakespeare: The invention of the human.* London: Fourth Estate.

Bowlby, J. (1971). *Attachment and loss, Volume 1. Attachment.* London: Hogarth Press (reprinted by Penguin, 1978).

Dunbar, R. I. M. (1996). *Grooming, gossip and the evolution of language.* London: Faber & Faber.

Erasmus, D. (1508). *Praise of folly* (Ed. & trans. R.M. Adams). New York: Norton.

Goldberg, S., Grusec, J. E., & Jenkins, J. M. (1999). Confidence in protection: Arguments for a narrow definition of attachment. *Journal of Family Psychology, 13,* 475-483.

Jacobson, R. (1988). "Linguistic and poetics;" and "The metaphoric and metonymic poles." In D. Lodge (Ed.), *Modern criticism and theory: A reader* (pp. 31-61). Longman: London.

James, H. (1884). The art of fiction. *Longman's Magazine,* September, Reprinted in *The Portable Henry James* (1951) (Ed. M.D. Zabel) New York: Viking, (pp. 391-418).

Jenkins, J. M., & Greenbaum, R. (1999). Intention and emotion in child psychopathology: Building cooperative plans. In P. D. Zelazo, J. W. Astington, & D. R. Olson (Eds.), *Developing theories of intention: Social understanding and self control* (pp. 269-291). Maywah, NJ:: Erlbaum.

Lovejoy, C. O. (1981). The origin of man. *Science, 211,* 341-350.

Neisser, U. (1963). The imitation of man by machine. *Science, 139,* 193-197.

Oatley, K. (1992). *Best laid schemes: The psychology of emotions.* New York, NY: Cambridge University Press.

Oatley, K. (1999). Why fiction may be twice as true as fact: Fiction as cognitive and emotional simulation. *Review of General Psychology, 3,* 101-117.

Oatley, K., & Jenkins, J. M. (in press). *The heart's reasons.*

Shakespeare, W. (1623a). *As you like it* (Ed. A. Brissenden). Oxford: Oxford University Press (1993).

Shakespeare, W. (1623b). *Hamlet* (Ed. J. Jenkins). London: Methuen (current edition 1981).

Shakespeare, W. (1623c). *Henry IV Part I* (Ed. D. Bevington). Oxford: Oxford University Press (1987).

Shakespeare, W. (1623d). *The Tempest* (Ed F. Kermode). London: Methuen.

Simon, H. A. (1967). Motivational and emotional controls of cognition. *Psychological Review, 74,* 29-39.

Wilson, E. O. (1998). *Consilience: The unity of knowledge.* New York: Knopf.

We Must RE-MEMBER to RE-FORMULATE: The M System

Doug Riecken
IBM Watson Research Center and Rutgers University
riecken@us.ibm.com riecken@home.com

1 Introduction

This paper provides introductory discussion on the M system. M is a study of an architecture that supports integrated multiple reasoning processes and representations. This architecture has been applied and evolved through a series of different domain problems:

1. Wolfgang, a system that learns to compose music (Riecken 1989, 1992a),
2. adaptive user interfaces (Riecken 1991a, 1991b, 1992b), and
3. the M system (Riecken 1994), a software program that acts as an assistant to a user by classifying and managing domain objects in a multimedia conferencing system.

The goal of this work is to develop a theory of mind that enables common sense reasoning to be applied in M. M was designed to observe situations and formulate beliefs about these situations regardless of the truth of the beliefs. It appears that humans observe and believe, and then over time continue to improve their knowledge of their beliefs while many types of computer programs just get stuck!

I take the position that common sense learning is a time variant problem and that learning is a constant series of viewing similar situations from different points of view over time. You can not learn something until you learn about "it" from several points of view. Good common sense reasoning and learning results from the ability to perform reformulation on an idea, concept, or problem. Reformulation requires rich fluid representations, multiple modalities of reasoning, and a range of experiences over time. Minsky's Society of Mind (SOM) Theory (Minsky 1985) is the essence of the study and implementation of M.

M was initially designed and implemented in my work at AT&T Bell Laboratories. M functioned as a software process that recognized and classified objects and actions performed by humans in a multimedia desktop conferencing system we developed at Bell Labs, called the Virtual Meeting Service. In this Computer Supported Cooperative Work (CSCW) service, participants worked in a Virtual Meeting Room (VMR) on a series of tasks. Each participant has a personal M program that watches all legal actions performed on all legal objects in the VMR by all participants. Each personal M attempts to recognize and classify what all the participants are doing and then support its respective user to recall items and actions that might relate to the user's current task and context.

2 VMR as a CSCW environment

To design a prototype model that can perform CSCW classifier functions, a specific CSCW environment was identified. Based on my previous work at Bell Laboratories in developing AT&T's Virtual Meeting Service, I defined the CSCW environment on the idea of a virtual meeting room (VMR) that supports multimedia desktop conferencing.

In a VMR, participants collaborate via computers and shared applications that provide users with documents, whiteboards, markers, erasers, staplers, copy machines, and many other such objects. The actual VMR is a complex set of data structures hosted on a server platform that maintains a consistent state view of a VMR session for all legal participants.

Conceptually, a VMR is a virtual place where one or more persons can work together even though the individuals are physically separated. An example of a VMR is a computer hosted place where individuals physically located in New Jersey and England can meet and work. In a VMR, the individuals share and create information in a variety of media ranging from text to images to drawings.

VMRs also support the functionality of persistence, thus VMRs can exist over arbitrarily long periods of time. A VMR is like a real meeting room where individuals can work, leave at the end of a day while leaving behind all documents and other objects, and then return at a later point in time to continue the work at hand.

3 The M system

The M system is a computer model (program) that performs classification tasks in a VMR. M is a system that applies "common sense" reasoning and knowledge to formulate classifications of VMR domain objects. M's reasoning does not rely on the content contained in VMR objects (e.g., documents), but instead M observes simple contextual cues and features present in typical VMR situations. Simply put, M reasons based on context, not content.

The power of M's "common sense" reasoning results from integrating "simple" facts and rules asserted from different lines of reasoning. M's model is a collection of simple facts and ideas about user collaboration in a VMR.

In order to develop a theory of common sense reasoning, I have studied and designed systems that support multi-reasoning processing. This appears to be essential in that the common sense "things" we understand as humans results from integrating many very simple, sometimes trivial, pieces of information about the world around us. A good theory on common sense reasoning might require that reasoning integrate information based on such distinct views as time, space, and function. To examine such a theory, we first must select a "world" in which to perform common sense reasoning.

The VMR world is a much simpler world than our own physical world. So, in order to better understand how to make use of many very simple facts, some which we use all the time without realizing, I have continued my study via the VMR world. The VMR world is an explicit finite problem space in which a formal representation of the useful information might provide a better understanding on how a system, biological or in silicon, might be able to reason about objects and actions within a VMR.

The M system is a multi-strategy classifier system architecture contains the following:

- semantic net functions
- rule-base system
- scripting system
- multi-ranked blackboard system based on Minsky's Trans-Frames in SOM

The design of M must enable M to function as a useful assistant to a human user. This implies that M's classification and knowledge of users working in a VMR must appear to a user to make sense from the user's point of view. Thus M must reason in a manner consistent with the user.

4 Function of M

In a VMR, each user is supported by a personalized M assistant and the VMR world is composed of domain objects (e.g., electronic documents, electronic ink, images, markers, white boards, copy machines, staplers, etc.) upon which users apply actions. The M assistant(s) attempt to recognize and define relationships between objects based on the actions applied by the users to the VMR world and the resulting new states of that world. For example, in a VMR world, there may exist a set of domain objects – such as several documents. Further, the VMR participants may apply actions to these documents such as annotating over them collectively or joining/appending them together. M attempts to identify all domain objects and classify relationships between subsets of objects based on their physical properties and user applied actions.

5 Simple example

A simple example would be 2 adjacent documents which a user then annotates by drawing a circle to enclose them together. Thus based collectively on (1) spatial reasoning of the nearness of the 2 documents and the circle, (2) structural and functional reasoning of the circle enclosing the 2 documents, and (3) casual reasoning of the semantic action of enclosing objects – M can infer and explain a plausible relationship between the 2 documents.

6 Organizing the VMR workspace

Consider a typical group of designers working in a brainstorming session held within a real physical room. By the end of such a working session, the designers will have created and used many documents, bullet lists, diagrams, notes, post-its, and other such items. Based on the properties of a physical room, the participants could organize themselves and the objects in the room using tables, walls, and whiteboards. Documents and other objects could be spatially organized and located for ease of access by the meeting participants. Typically, the designers would be able to view, engage, review, and reformulate various conceptual relationships over all the physical materials and information generated as the meeting progressed.

When we port the designer's brainstorming session to a VMR, their view of the work environment is significantly constrained to the physical size of their respective computer screens (e.g., ~ 1000x1000 pixels at best). What if M took on the responsibility to organize the output and interactions of all the participants? In essence, M assists a user to access and manipulate many different materials created and used during a meeting, independent of where the materials are located within a VMR or when the materials were last used or created.

M can generate and present various classifications representing conceptual views of VMR objects created and used by the participants. Thus, each participant can ask, via dialog boxes or direct manipulation techniques, their respective M assistant to present organized views of the various related materials used during a meeting.

Functionally, M observes the actions performed by VMR participants and attempts to reason how the current actions applied to VMR objects relate to other VMR objects and previous actions. As a participant interacts with an object, such as a document, M can provide the user with contextual hyperlinks to related objects, such as documents, drawings, notes, lists, post-its, and pen annotations. One of M's fundamental responsibilities is to assist a user to (RE)formulate relationships between all objects in a VMR.

Specifically, M attempts to maintain simultaneous theories of how objects in a VMR might relate. This enables M to provide participants with multiple views or access of related materials – thus, M and a user can reformulate the relationships between VMR objects.

While M maintains an extensive schema for organizing and representing a VMR, it must also allow the user to (RE)define existing and new relationships and hyperlinks within this schema. This safe-guards that M never takes control away from the user.

A useful idea in building a mind is the application of set theory and partial orderings as clever tricks to think about. Minsky's K-lines in SOM (Minsky 1980, 1985) are extensive sets of partial orderings of the enormous number of "facts and rules" that worked in previous situations and life experiences. The trick in these various learned ASSOCIATIONS is that they are members of various sets representing some learned idea, fact, concept, or process. Marvin has a wonderful play with words to remind us of this powerful idea. In Society of Mind, he writes the word remember as RE-MEMBER. We RE-MEMBER by using some members of a set of members that worked in some previous situation.

7 Design of M

The design of the M system required a formal world representation of a VMR. The world definition contained knowledge about all domain objects and the legal actions which can be applied within the world; the legal set of VMR situations.

The design goal of the M system was to recognize and classify actions and objects in a VMR world based on a "common sense" reasoning approach, instead of relying on "understanding" the content of the VMR objects via some form of natural language processing. In defining the ontology of M's knowledge-base, the following two tasks were required:

- develop a theory of the VMR recognition and classification process
- formulate a representation of the problem domain for all domain objects and actions

AI research has identified problem solving methods for ill-structured problems (Newell 1969, Simon 1973) were a set of heuristic processes generate a solution over a defined problem space. M's "common sense" reasoning relies on heuristics as it observes the world and applies contextual, not contentual, information about the objects and actions relating to a VMR situation. The design theory of M required a multi-strategy reasoning approach.

In a VMR situation, there are many simple and sometimes obvious cues which when combined together formulate a plausible theory of how objects relate. M integrates different reasoning processes which assert very simple facts into shared data structures representing the generation of a classification theory for a VMR situation. Presently, the M system is designed with five modal reasoning processes which collaborate to develop classification theories. The modalities of reasoning are: structural, functional, spatial, temporal, and causal.

8 M's recognition and classification process

M examines a VMR situation via the collaboration of distinct reasoning processes. The design theory for M partitions the problem solving process, the classification of VMR situations, into the following ordered sequence of functional tasks:

- represent a VMR situation consisting of an action, the pre VMR state prior to the action, and the post VMR state resulting from the action
- identify and characterize the object(s) involved in an action – this requires enumerating all known properties of the object(s)
- propagate the constraints relating to the object(s) and action to all reasoning processes responsible to classify the VMR situation
- have the reasoning processes collaborate to develop potential classification theories of the VMR situation
- restrict the range of plausible theories in order to avoid combinatoric growth

9 Ms' architecture

M's architecture consists of the following five key components representing knowledge of domain objects, legal actions, and legal situations: (1) a semantic net system, (2) a rule-based system, (3) a scripting system, (4) five distinct reasoning processes (inference engines) and (5) a blackboard system consisting of an ordered set of blackboards.

10 SEMANTIC NET SYSTEM

The semantic net system is implemented as a spreading activation network over sets of qualifiers (e.g., size, position, color, etc.) which collectively represent domain object characteristics. These qualifiers represent the facts associated with an applied action denoted in an input record. Each qualifier acts as a state machine representing the current legal property value of a VMR object. For example, the color qualifier can enter into a state representing the color of an object or a shape qualifier can enter into a state representing such shapes as square, circle, etc. The basic idea is this – when an object is identified via the I/O system, the corresponding qualifiers within the semantic net collectively become active representing the correct property states of the respective object. As these qualifiers become active in a specific state, they become facts which are asserted to M's rule-based system.

11 RULE-BASED SYSTEM

M's rule-based system performs several important functions. As facts (in the semantic net) are asserted, they in turn satisfy specific pre-conditions expressed in the antecedent of given rules. Thus, as the antecedents of such rules evaluate as true, this enables the consequence of each respective rule to be asserted. This can have the following two results. First, new facts expressed in a rule's consequence are asserted respectively to the semantic net; this then can have an iterative effect over the firing of new rules and the instantiation of other facts. Second, as new rules fire and new facts are instantiated, M's reasoning processes can in turn apply this new information to strengthen or weaken or create or purge the various theories representing a VMR world.

As various facts and rules evaluate as true, this directly influences M's scripting system and reasoning processes as they evaluate and apply various scripts of partial plans provided by the scripting system. In essence, we can view M's rule-based system as a collection of domain conditions that when satisfied are applied to bias the selection of partial plans from M's scripting system by M's reasoning processes to create and "explain" relationships between VMR objects.

12 SCRIPTING SYSTEM

M's scripting system is a corpus of partial plans that have demonstrated frequent success in previous classification problems. In M, a script is a partial ordering of elements in a set; the set represents an interval of time during which a consistent pattern of facts and rules have frequently been applied successfully to predict the state of some object(s) following some action. M's design of a script is based on Schank and Abelson's presentation of scripts (Schank 1977).

An important feature of M's scripting system entails the use of coefficients to weight each script's potential to either initiate or improve upon a theory which attempts to classify and represent some set of actions, objects, and relationships within a VMR. Functionally, these weighted scripts bias the various reasoning processes to dynamically rank all coexisting theories where each theory is formulated on one of the individual blackboards. These weighted scripts serve to minimize combinatoric growth of all possible classification theories. The reasoning processes will select weighted scripts that formulate or improve only the top seven ranked theories.

13 Multi-strategy reasoning

M's architectural design was based on a theory of integrated reasoning processes; sometimes referred to as integrated "agents" or inference engines. This multi-strategy reasoning ability of M allows the system to formulate different points of view while performing recognition and classification tasks.

In the applied domain of the VMR, it was useful and typically necessary that M simultaneously derive and manage several theories representing the actions of VMR participants and the state of all VMR objects (e.g., documents, files, pens, markers, erasers, etc.). This was due to the fact that certain classifications were not immediately obvious – either (1) they emerged over time or (2) given contextual situations enforced reformulation of existing classifications.

In my study, one of the key research issues concerned the management of the different reasoning processes as they collectively formulated multiple theories to recognize and classify a VMR world. This management function required a technique for the processes to "communicate" and leverage key information relative to distinct simultaneous classification theories of a given VMR situation.

In developing a design theory of M as an architecture of integrated reasoners, it was desirable to define a framework in which simultaneous theories of a world could be dynamically generated, ranked, and modified. For the applied problem of the VMR world, five different reasoning processes were required and implemented as distinct inference engines. The five types of reasoning supported in M are:

- structural
- functional
- spatial
- temporal
- causal

The integration and management of these inference engines was achieved via a traditional shared data structure and governing processes known as a blackboard system. In the M system, each reasoning process served as a knowledge source (KS) which inter-worked with other KSs via the blackboard system.

The design and implementation of M's blackboard system resulted in two unique features. First, M consisted of a dynamically ordered set of blackboards. Each blackboard hosted a distinct theory representing M's recognition and classification of a VMR situation. The set of blackboards were ranked based on the strength of each theory's probable correctness. Second, the structure for representing information posted by KSs to a given blackboard was based on Minsky's Society of Mind Transframe.

14 Blackboard systems

Blackboard systems are a means of implementing dynamic, opportunistic behavior among cooperating reasoning processes that share intermediate results of their efforts by means of a global data structure (the blackboard). Penny Nii (Nii 1989) describes the basic structure of a blackboard system in terms of three components:

- The knowledge sources (KSs). The knowledge needed to solve the problem is partitioned into knowledge sources, which are kept as independent processes.
- The blackboard data structure. The problem-solving state data (objects from the solution space) are kept in a global data store, the blackboard. KSs produce changes to the blackboard which lead incrementally to a solution to the problem. Communication and interaction among the KSs take place solely through the blackboard.
- Control. What KS(s) to apply when and to what part of the blackboard are problems addressed in control. Typically, a scheduling process performs the control function.

In addition to the organizational requirements, a particular reasoning (computational) behavior is associated with blackboard systems. The solution to a problem is built incrementally over time. At each control cycle, any reasoning assertion (e.g., data driven, goal driven, forward chaining, backward chaining, etc.) can be used. The part of the emerging solution to be attended to next can also be selected at each control cycle. As a result, the selection and the assertion by KSs are dynamic and opportunistic rather than fixed and preprogrammed.

15 Ranked blackboards

M's blackboard system consists of a dynamic set of ranked blackboards which are allocated and reallocated as needed. The maximum number of blackboards allocated at any given moment is seven. Each blackboard contains an emerging classification theory over some subset of actions and objects. Basically, an emerging theory can be thought of as a hypothesis to be proved by M's reasoners. M's reasoners attempt to develop a strong theory by individually applying axioms to a given theory's hypothesis on a blackboard.

As M observes actions being performed by VMR participants, M's semantic net, rule based system, and scripting system assert new facts, rules, and scripts respectively via the five KSs. The KSs collaborate by applying this information as axioms to the respective blackboard of a given classification theory. Further, as M computes the weighted scripts for each blackboard, the theories with the greatest weighted sum are ranked high to low, thus defining the dynamic ordering of blackboards.

16 Trans-frames

When a KS posts an axiom to a blackboard, this information can be viewed either as some type of modal information reflecting a modality of reasoning (e.g., spatial, temporal, structure, etc.) and/or some set of "conceptual dependency information" representing an action. The fundamental data structure of an individual blackboard is based on Minsky's Trans-frame. The Trans-frame provides a representation of an action, a trajectory between two situations; this information represents the pre and post states of a VMR situation.

The "conceptual dependency information" depicted in a Trans-frame structure includes:

- the actor performing the action
- instrument used by actor to perform action
- the action applied to some object(s)
- the object(s) with pre state properties
- the object(s) with post state properties
- the difference(s) between the pre and post properties
- list of plausible goals addressed by the action
- causal effect of the action

The Trans-frame structure provides a canonical form which enables M to effectively compare:

- different theories or sub-theories posted over the ranked blackboards,
- the various weighted scripts contained within the scripting system with a given theory posted on a blackboard, and
- the pre and post properties of the object(s).

Embedded within a Trans-frame structure are two object property graphs representing the object(s) pre and post state properties. This graph-based structure represents an object's properties based on the different modalities of reasoning. The application of this structure was reported by Winston et al. (Winston 1983) and Mitchell et al. (Mitchell 1986). The object property graph depicts properties based on their functional, structural, spatial, and temporal values and enables inference across different modal reasoning. Like the Trans-frame, the object property graph is a canonical form which enables effective evaluation and comparison of multiple objects.

References

M. Minsky. K-lines: A theory of memory. *Cognitive Science*, 4,:117–133, 1980.

M. Minsky. *The Society of Mind.* Simon and Schuster, New York, 1985.

T.M. Mitchell, R.M. Keller, and S.T. Kedar-Cabelli. Explanation-based generalization: A unifying view. In R. Michalski, editor, *Machine Learning, Volume 1, Number 1,*. Kluwer Academic Publishers, Netherlands., 1986.

A. Newell. Heuristic programming: ill-structured problems. In J. Aronofsky, editor, *Progress in Operations Research, Volume 3,*. Wiley, New York, 1969.

H.P. Nii. Introduction. In V. Jagannathan, R. Dodhiawala, and L. Baum, editors, *Blackboard Architectures and Applications,*. Academic Press, New York, 1989.

D. Riecken. Goal formulation with emotional constraints: Musical composition by emotional computation. In *AAAI Proceedings First Annual Conference on Innovative Applications of Artificial Intelligence, Stanford University,*, Cambridge, Massachusetts, 1989. AAAI/MIT Press.

D. Riecken. Adaptive direct manipulation. In *Proceedings of IEEE International Conference of Systems, Man and Cybernetics*, pages 1115–1120,, Charlottesville, VA., 1991a.

D. Riecken. Auditory adaptation. In *Proceedings of IEEE International Conference of Systems, Man and Cybernetics*, pages 1121–1126, Charlottesville, VA., 1991b.

D. Riecken. Human-machine interaction and perception. In M. Blattner and R. Dannenberg, editors, *Multimedia Interface Design.* ACM Press, New York, 1992a.

D. Riecken. WOLFGANG - A system using emoting potentials to manage musical design. In M. Balaban, K. Ebcioglu, and O. Laske, editors, *Understanding Music with AI: Perspectives on Music Cognition.* AAAI/MIT Press, Cambridge, MA, 1992b.

D. Riecken. M: An architecture of integrated agents. *Communications of the ACM*, 37(7), July 1994.

R.C. Schank and R.P. Abelson. *Scripts, Plans, Goals, and Understanding*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1977.

H. Simon. The structure of ill-structured problems. *Artificial Intelligence*, 9, 1973.

P.H. Winston, T.O. Ginford, B. Katz, , and M. Lowry. Learning physical descriptions from functional definitions, examples, and precedents. In *Proceedings National Conference on Artificial Intelligence*, pages 433–439, Cambridge, MA., 1983. AAAI/MIT Press.

Behavioral States: Linking Functional and Physical Descriptions

Matthias Scheutz
Department of Computer Science and Engineering
University of Notre Dame, Notre Dame, IN 46556, USA
mscheutz@cse.nd.edu

Abstract

In this paper, I will introduce the notion of "behavioral state" as a means to bridge the gap between functional specifications and their implementations, borrowing from ethology and to some extent from research in behavior-based robotics. First, I will briefly sketch some of the problems resulting from mere functional descriptions for the designer of a mind. Then I will define the notion of behavioral state and locate its place as mediator between functional and physical states. After sketching the way in which I forsee behavioral states could be used in the design of minds, I will conclude that an intermediary level of architectural specification between functional and physical description will be of great advantage (if not necessity) in designing a mind, regardless of the success of the notion of behavioral state.

1 Introduction

Ever since Descartes, philosophers interested in the mind have divided the world into the mental and the physical realm contemplating exactly how these two realms are related. While this issue is still far from being resolved, today's most commonly held view on the "mind-body" problem in the philosophy of mind is *functionalism*, i.e., the claim that mental states *are* functional states, which somehow "supervene" on physical states.[1] The general understanding is that mental states (i.e., states such as "believing that *p*" or "desiring *x*", or even psychological predicates such as "pain" or "pleasure") can be explained in terms of functional states and functional architectures.

Besides the fact that to my knowledge no one has attempted to account for concepts from folk psychology by specifying *in detail* a functional architecture for them, it seems to me that there will still be major obstacles for the artificial intelligence researcher who wants to build actual agents (that realize a given functional architecture). Even if such an architecture could be provided the question remains how functional states are related to physical states? Furthermore, should a combination of "computational-physical" states be used as realizers instead of physical states alone? What constraints does the architecture impose on the implementing system?

Relating functional states *directly* to physical states is very unlikely to succeed in the light of multiple realization arguments for functional architectures (the more complex the architecture gets, the less we will be able to see what kinds of possibly very diverse physical systems will share the functional specification). The level of functional specification of the psychology of minds will be too high and abstract a level of description to suggest *possible implementions* of the functional states (not to mention all the problems connected with the involved notion of "implementation" or "realization" that seem to be largely ignored by the philosophical community).[2]

It is my conviction that functional specifications of psychologies are not sufficient to suggest ways to build a mind. To be of any practical importance in designing a mind at all, a level of description of a cognitive architecture has to incorporate at least *some* of the relevant physical properties of its possible implementations, which will constrain both possible implementations as well as functional architectures. In this article, I will suggest such an intermediary level, which I call *the level of behavioral states*. This level of description is largely inspired by ethological studies of animal behavior and to some extent by research in behavior-

[1] The questions of exactly how these states supervene on the physical and in what kinds of structures they are realized are rarely addressed in detail, let alone answered satisfactorily. This is most likely due to the fact that the notions of "realization" and "supervenience" are mostly used as unexplained "primitive" terms in the philosophical literature (which is quite surprising given the theoretical importance and practical consequences that hinge upon them). Although some have attempted more or less precise definitions of "realization"—e.g., Kim, Block, et al.—these definitions are not very helpful for those who, interested in building minds, are trying to understand the relation between architectures and their implementations.

[2] Note that this obviously does not hold for all functional specifications: a functional specification of an abstract finite state automaton, for example, can be easily related to physical states in a standard PC by "implementing" the automaton in a programming language.

based robotics and will therefore bear the signia of these intellectual sources very visibly on its sleeves.

First, I will briefly point to problems resulting from mere functional descriptions for the designer of a mind. Then I will introduce the notion of "behavioral state" and locate its place as mediator between functional and physical states. I will sketch the way I forsee that behavioral states could be used in the design of minds on a simple cognitive architecture. Finally, I shall conclude that regardless of the success of the notion of behavioral state in designing minds, an intermediary level of architectural specification between functional and physical description will be of advantage (if not necessity) in designing a mind.

2 Functionalism

2.1 The Functionalist Picture

A functional specification of a cognitive architecture consists a set of input states, a set of output states, and a set of "inner" or "functional" states together with a specification of how they are causally related. That way it is possible to determine what state a cognitive system will be in next, given the current state and all the input conditions.[3] While input and output conditions have to be tied to physical inputs and outputs, the functional states do not require a direct correspondence to their physical realizers as expressed in the phrase that "functional states supervene on physical states" (e.g., see Kim, 1997). This lack of a "direct" correspondence between functional and physical states is what gives functionalism its explanatory power, while keeping it metaphysically palatable: it combines advantages of behavioristic approaches to mind (i.e., considering solely the input-output behavior of an organism) with advantages of identity theories (i.e., mental state/event tokens are physical state/event tokens) leaving out the pitfalls of both such as the lack of being able to account for "inner states" in the former, and the requirement of type identities between mental and physical state/event types of the latter. Yet, this strength comes at a price: it is not clear what it means to *implement* or *realize* a functional architecture (see Scheutz, 2000a).

2.2 Implementation of a Functional Architecture

So what are the implementation conditions for a functional architecture? To say that a system implements a functionalist description is to require that in addition to the input and output mapping, it has to get the mapping of the inner states right. Usually, these "inner states" are assumed to be multiply realizable, therefore the mapping has to be a many-to-one mapping from physical states to functional states (very much in the spirit of Chalmers, 1997). Yet, inner states are viewed by functionalists as intrinsically relational states, being *mutually defined* by all states in the functional architecture.

To illustrate this interdependence, consider, for example, the following automaton, which has two inner states 'E' and 'O' standing for "even" and "odd". Depending on whether the number of '1's that the automaton has seen so far is even or odd, it outputs either 'a' or 'b', respectively.

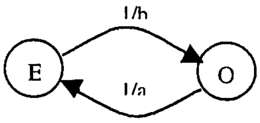


Figure 1: The even-odd transducer with two inner states.

A functionalist account (e.g., see Block, 1996) of what it means to be in state E would look like this:

> Being in E $=_{\text{def}}$ Being an x such that $\exists P \, \exists Q$ [x is in $P \wedge$ (if x is in P and receives input '1', then it goes into Q and outputs 'b') $\wedge$ (if x is in Q and gets input '1', then it goes into P and outputs 'a')].[4]

Since it is only claimed that there has to be an arrangement of physical states that corresponds to the functional states in a way that preserves inputs and outputs as well as transitions between states, it is possible for one physical state to serve as the instantiation of more than one functional state (and vice versa). Therefore, the correspondence between physical and functional states is not necessarily that of a mapping between physical types and functional types (let alone a 1-1 mapping), but rather that of a relation that preserves state transitions. "Implementation of a functional architecture", therefore, has to be viewed as some sort of "bisimilarity" between functional and physical architecture rather than some sort of *isomorphic* relation from a functionalist point of view.[5] As a consequence, finding a relation between a given functional architecture and a set of physical states together

[3] Of course, the behavior elicited by the organism realizing the cognitive system is specified as well.

[4] Note that the existential quantifiers could be viewed as ranging over properties or as picking out particular physical states of the system.

[5] The notion of "bisimilarity" is defined as follows: let I and O be two finite sets (e.g., the sets of input and output states, respectively) and let $M_1=\langle S_1, \rightarrow_1 \rangle$ and $M_2=\langle S_2, \rightarrow_2 \rangle$ be two structures with domains S_1 and S_2, respectively, where relation $\rightarrow_1$ is defined over $S_1 \times I \times S_1 \times O$ and relation $\rightarrow_2$ is defined over $S_2 \times I \times S_2 \times O$. These structures are then said to be *bisimilar* if there exists a non-empty relation R between S_1 to S_2 such that for all $s_1 \in S_1$, $s_2 \in S_2$, $i \in I$, and $o \in O$ the following two conditions hold: (1) i f $R(s_1, s_2)$ and $(s_1, i) \rightarrow_1 (t_1, o)$, then $(s_2, i) \rightarrow_2 (t_2, o)$ and $R(t_1, t_2)$, and (2) if $R(s_1, s_2)$ and $(s_2, i) \rightarrow_2 (t_2, o)$, then $(s_1, i) \rightarrow_1 (t_1, o)$ and $R(t_1, t_2)$.

with their causal transitions will be an intractable problem for reasonably large sets of states.[6] In other words, a mere functional specification of a cognitive architecture is not going to be of any help in designing a realizer.

3 Behavioral States

3.1 An Ethological Perspective

To overcome the difficulties of tying functional specifications to physical implementations, I suggest to consider work done in animal behavior research as a venture point. According to animal behaviorists (e.g., McFarland, 1981), animal behavior can be categorized in terms of

(1) reflexes (i.e., rapid, involuntary responses to environmental stimuli)
(2) taxes (i.e., responses orienting the animal towards or away from a stimulus)
(3) fixed-action patterns (i.e., time-extended sequences of simple responses)

While (1) and (2) are solely connected to external stimulation, (3) can have a contributing "internal" component as well (fixed action patterns can be "motivated"; take, for example, the "egg-retrieving" behavior of the greyling goose, see Lorenz, 1981, or Lorenz and Leyhausen, 1973). All three kinds of behaviors can be combined in complex ways to form hierarchies of behaviors (see figure 2).

In these behavioral structures, behaviors form "competitive clusters", in which behaviors are mutually exclusive (e.g., in figure 2 the "fighting behavior" is such a competitive cluster comprising the mutually exclusive behaviors "chasing", "biting", and "display").

To make these ideas of behavioral hierarchies more concrete, I will introduce the notion of *behavioral state*, which roughly corresponds to what is indicated by a "circle" in figure 2. Putting it crudely, a behavioral state is a state an individual is in if it performs a particular behavior (e.g., such as "food handling" or "looking out for prey").[7] "Behavior" is meant be understood in a *wide sense* to include behaviors that are not necessarily observable from the outside alone (such as "memory recall" or "thinking", in general). Hence behavioral states are not simply combined input-output states, but rather they are some sort of "inner states" of an organism, states in which the organism is if it performs a particular kind of behavior. Note, however, that nothing is implied or claimed about a particular physical correlate of a behavioral state—it might or might not exist (I will return to this issue later).

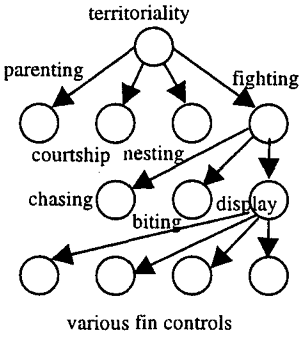


Figure 2: A part of a behavioral hierarchy for the male stickleback fish (see Lorenz, 1981). The various fin controls can be divided further into rays of each fin, the muscle fibers for each ray, and the motor neurons for each fiber.

Behavioral states are not restricted to "motor actions", but include sensory actions as well as more abstract proprioceptive and reflective actions (such as monitoring inner physiological states, generating images, producing plans, recalling poems, analyzing pictures, making logical derivations, etc.). The latter ones are more "abstract behaviors", which are mostly (if not completely) internalized and often involve solely parts of the cognitive architecture; in fact, they might not result in any externally observable change at all (a mathematician contemplating abstract objects and manipulating their representations in her mind might not need any stimulation from the outside world in performing this task, nor might any motor action result from it—this "brain in a vat"-idea with sustained cognitive activity whilst lacking external interaction seems to be at least conceivable in principle).

Memory and reflective processes, for example, are then viewed as special kinds of behavirol processes that lead to actions performed directly on the cognitive architecture, as opposed to the effectors of the individual.

In general, an individual will be in many behavioral states at the same time reflecting the fact that (1) some behaviors are contained in or shared among others (for example, searching for food as well as searching for a mate will both involve locomotion, despite the fact that the kind of search might be different), and (2) that many behaviors are performed in parallel (such as monitoring my hand as I move it to pick up an object).

[6] This problem will indeed be at least as hard as "graph isomorphism", which itself is believed to be in NP proper.

[7] A note of terminology: while it is common usage to use "mental states" and "functional states" to refer to states of an individual's mind, the notion of state is not exclusively used to describe "static" entities, but often times serves the role of a general term that subsumes *states* as well as *events*, i.e., processes. In a sense, the term "behavioral state" should have been avoided in favor of "behavioral processes", as the latter emphasizes the dynamic character of the activity taking place in the individual. Following established terminology, however, I will continue using the term "behavioral state", even if (systematic) dynamic changes in the individual are being referred to.

3.2 Behavioral Architectures

In a sense, the classical ethological picture outlined above is mainly concerned with the relation between various behaviors, it only depicts some causal relations between behaviors, and is, therefore, a functional specification of the behavioral architecture. Yet, partly implicit in and partly external to this picture is information about the time constraints as well as the strength of interactions and influences among behaviors (as studied and gathered by animal behaviorists). In other words, the picture is *incomplete* in so far as it leaves out essential implementation details that cannot be retrieved from a picture like figure 1 alone. Without these implementation details, however, some behaviors would not be the behaviors they are, since what distinguishes them from other behaviors might just be constraints on timing and strength of response (take, for example, a retraction reflex caused by touching a hot plate with your finger as opposed to the same movement being performed very slowly). Furthermore, the strength and configuration of interactions between behaviors is an integral part of their defining characteristics, which cannot be captured by a causal structure alone: suppose behavior A *causes* behavior B. Then this can happen in many different behavioral arrangments, for example, by A enforcing B directly or A suppressing C, which in turn inhibits B, or by A enforcing D, which enforces C, etc. Implicit in A (as defined by an animal behaviorist, say) is already information, which of these possible arrangments are realized in the animal. Hence, the causal structure might get restricted by the behavioral structure if (some of) the information implicit in the definition of behaviors is made explicit. In the following, I will briefly sketch how behavioral states can be defined to explicitly incorporate some of the otherwise implicit aspects of behaviors.

3.3 The Structure of Behavioral States and Networks

First and foremost, each behavioral state has an *activation level* and a *behavior* associated with it. This activation depends on various factors: (1) its own activation level, (2) the activation level of other states, (3) possible inputs from exteroceptive and proprioceptive sensors, (4) energy constraints, and (5) decay over time.[8] The behavior associated with a behavioral states can either be a simple behavior (such as reflexes and taxes), or either a more complex fixed behavior (such as fixed action patterns) or a more complex adaptive behavior (which results from the interplay of fixed action patterns, reflexes, and taxes). The term "adaptive" indicates that the latter kinds of behaviors can change over time, i.e., they can be learned, altered, etc. (utilizing the dynamic interplay of behavioral states).

Behavioral states are connected via inhibitory and excitatory links to other behavioral states and possibly to sensors (via "information channels", i.e., filtering mechanisms that select parts of one or more sensory inputs and combine them in particular task-specific ways). Connections between behavioral states have a distance associated with them (expressed in terms of a time-lag), reflecting the "distance in space" that a signal has to travel from one locus of action to interact with another, allowing temporal as well as spatial integration of incoming signals.

Groups of behavioral states that are connected via mutually inhibitory links form so-called "competitive clusters". They inhibit each other to various degrees, while usually entertaining excitatory connections to lower and upper level states (and possibly to some behavioral states of other clusters at the same level as well). In such a cluster the behavior associated with the state with the highest activation is activated and all behaviors of the other states are suppressed.[9] This way hierarchical structures similar to the one in figure 1 can be defined which reflect the relationship between behaviors and in part also the complexity of each behavior associated with the various states (the lowest levels corresponding to simple reflex-like, reactive behaviors—this level has been explored in great detail in behavior-based robotics, e.g., see Arkin, 1992, or Brooks, 1986).

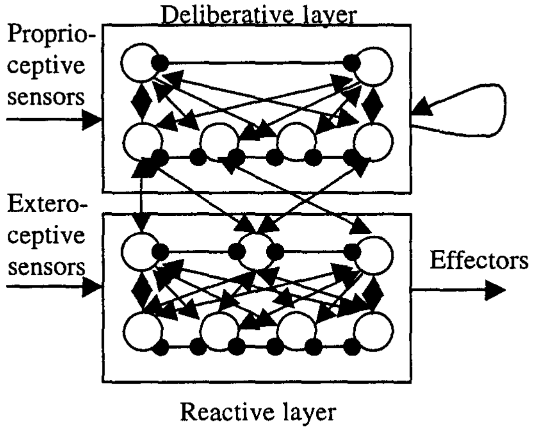


Figure 3: A hierarchy of behavioral states viewed as a two-layered architecture consisting of a deliberative and a reactive layer. Links with arrows indicate excitatory connections, links with circles inhibitory ones. The behavioral units in the deliberative layer do not operate on effectors, but perform internal operations (such as

[8] I will not be able to address issues related to last two points in this paper.

[9] There is evidence that similar mechanisms are at work in animals that inhibit all behaviors with lower activation values, e.g. see Lorenz (1981).

memory lookups, symbolic combinations, etc.).

With respect to the spread of activation, networks of behavioral states are very similar to I(interactive) A(ctivation) and C(ompetition) networks (e.g., see Rumelhart and McClelland, 1986). Therefore, results from connectionist research about effects such as "blocking", "settling", "oscillation", "hysteresis", and others (often) apply *mutatis mutandis* to behavioral networks as well. The essential difference between IAC networks and behavioral networks is that the behavior associated with a behavioral state could affect via environmental feedback the activation level of the very state itself as well as the activations of other states. For example, a behavioral node representing the "search for black objects in visual field"-behavior might initiate occular motor commands that lead to the detection of a small black object by another node, which in turn inhibits the search node, thus decreasing its activation, which in a mere IAC network would have otherwise not decreased.

As already mentioned, not all behaviors will involve physical effectors; in fact, only low level behaviors will directly exert influence on them (these are behaviors that would normally be localized in what roboticists refer to as "reactive layer"). Higher level behavioral states will mostly operate on structures internal to the cognitive system (these states would be situated in the "deliberative layer"). For example, a "retrieve image of mother" node (assuming for a moment there is such a node), might initiate a search in long-term memory (possibly involving other behavioral states) for a particular image that is associated with the individual's mother. Or a "project-hand-move-forward" node might initiate a "simulated" hand movement in an emulator circuit, which is used to plan motions, resulting in a change in the circuit and as a consequence in other behavioral nodes (such as "collision detectors" in the emulator circuit, etc.).[10] A behavioral network divided into a layered structure consisting of a reactive and a deliberative layer is schematically depicted in figure 3.

There are special cases of behavioral states that do not have any behavior directly associated with them. Instead of initiating an action directly, they contribute to behaviors indirectly by influencing other behavioral states, and can, therefore, assume the role of affective states. A state corresponding to "hunger", for example, might receive inputs from proprioceptive sensors (i.e., a sensor monitoring the blood sugar or, more generally, the engery level) and exert positive influence on other states such as "search-for-food" (e.g., see Scheutz, 2000b). That way it is possible to entertain states that do not directly and immediately "cause" the individual to act in a particular way, but might have indirect, long-term effects on the individual (e.g., depression, memory loss, etc.).[11]

4 The Case for an Intermediate Level

4.1 The Relations between Physical, Functional and Behavioral States

So far, I have not explicated how physical and functional states relate to behavioural states as defined above. From an implementation perspective, behavioral states can be realized in many ways in different physical substrates. In brains, for example, they could correspond to a single neuron or to a group of neurons. They could be realized solely neuronally or maybe by involving other systems (such as the hormonal system) as well. Another physical medium, in which behavioral states can be realized, is the silicone of computers: computers can implement behavioral states by virtue of computational processes.

Some behavioral states might be (directly) "implemented" in the system in the sense that there exists a corresponding physical state or a set/sequence of physical states that are in *type correspondence* with the behavioral state. Other behavioral states might "supervene" on physical states in that there does not exist such a type correspondence—note that programs running on modern operating systems with virtual memory architectures exhibit such supervenience relations: when a program does not entirely fit into physical memory, it is loaded in parts on an "as-needed" basis, where different virtual memory locations get mapped onto the same physical memory location.

Another possibility for behavioral states to have no *fixed correlate* at all is to be only *partially* implemented (see Sloman, 1998) or to depend on environmental conditions (e.g., in terms of other behavioral states and/or environmental states—an example might be my performing the multiplication algorithm using paper and pencil: I am in a behavioral state which is implemented by a number of other states such as states of the paper and pencil, several visual routines, rule-retrieving memory processes and rule-following routines, etc.).

Behavioral states implemented in (sequences of) physical states are tightly coupled to their physical realizers (still allowing for multiple realizations), while behavioral states supervening on physical states do not exhibit such a coupling at all. They are realized by

[10] I am currently investigating various possibilities of implementing simple emulator circuits in terms of behavioral states.

[11] Compare this to standard philosophical talk about "pain *causing* wincing and groaning, etc.", where it is never clear whether pain always causes all the behaviors, exactly when the effects surface, whether showing the effects is necessary and/or sufficient for the individual to have pain, etc.

some physical states, but they might not show any systematic correlation to their realizers. For example, consider two networks of behavioral states, which are *functionally* identical except for the fact that the first explicitly implements a higher level behavioral state called "avoid-obstacle", which is active if the agent is engaged in obstacle avoidance behavior. The second one does not have such as state, but can still control the same obstacle-avoidance behavior. In this case, the behavioral state "obstacle-avoidance" has a physical correlate in the former and no fixed physical correlate in the latter (what corresponds physically to the "obstacle-avoidance" state in the latter is a complex sequence of patterns that might, under different circumstances, not correspond to this state at all, e.g., if the agent follows another agent, which is avoiding obstacles, and thus is a "follow other agent" state, which by pure chance causes it to go through the same sequence of physical states... see also Pfeiffer and Scheier, 1999, ch. 12 for another example).[12]

This aspect of behavioral states seems very similar to the kinds of functional states about which philosophers tend to worry, and maybe most of the "high-level" functional states such as "belief states", etc. are not directly (i.e., physically) implemented in the system (often the temr "emergent" is used in this context). Even so, these kinds of behavioral states still retain one aspect lost in the mere "causation talk" of functional architectures, and that is *time*!

4.2 Causation and Time

It has been pointed out by philosophers (e.g., see Chalmers, 1997) that there is an essential difference between functional descriptions of physical systems like clocks, combustion engines, CD players, etc. and the functionalist descriptions of minds: in the former case some aspects of the physical structure matter, they are essential to any system realizing the functional architecture. Thus, these physical aspects are (if not explicitly, so then implicitly) retained in the functional architecture, thereby constraining the set of possible realizers. In the latter case, however, it is the very functional structure itself—so it is claimed—that matters, that is, the patterns of causal organization regardless of the underlying physical structure. Therefore, only causal organization, or put differently, "the flow of causation" is retained in functionalist abstractions from the physical as *the essential aspect* of minds. But is this really true?

Real minds are intrinsically tied to their environments and thus affected by the temporal structures imposed on them. Timing plays a crucial role in every aspect of a cognitive architecture pertaining to the proper functioning and survival of the organism. Many recent studies in cognitive science emphasize the importance of time as opposed to "mere temporal order" (see, for example, Port and van Gelder, 1995).

What distinguishes *time* from mere (temporal) *order* (as implicitly provided by the notion of causality) is that in addition to order a *metric* is defined (on the set of time points), that is, a notion of *distance* in time. This notion of distance in time allows one to differentiate functions according to their temporal behavior that would otherwise be indistinguishable. Take, for example, two microprocessors that work at different clock speeds—functionally they are identical, yet there is an essential difference between them, which is usually also reflected on any price tag put on them: their speed (another example of a function, where time is the distinctive factor, would be vowel production and recognition).

Is it problematic that causation alone does not suffice to capture the temporal structure of cognitive architectures? I would claim: Yes. Imagine two different physical systems that share the same functional specification of a human mind, one a regular human, another the People's Republic of China "implementing the human brain" at a much, much slower pace (to use Block's example). A human body controlled by the People's Republic of China would fail terribly in the real world, because it could not react to its environment in due time.[13] Well, one might say, it would do just fine if everything surrounding it, that is, its environment had been "slowed down" appropriately. This objection, however, strikes me as severly flawed, since it would entail *a completely new physics* (as in our physical universe certain processes have to happen at a certain speed otherwise they would not be the kinds of processes they are). Whether a "slowed down version" of a human mind could control a "slowed down version" in such a "slowed down universe" (with possibly completely different physical properties) seems too speculative a question to be taken seriously. What seems to be a productive approach, however, is to ask whether it is possible to understand a certain architecture (that evolved or was designed to meet the temporal constraints of its environment) at a mere causal level? I suspect that the answer would be *no* for systems that are sufficiently complex (like brains of vertebrates or VLSI microchips, for that matter).

If, on the other hand, causal structure were augmented by temporal constraints (i.e., information about

[12] Note that it should be possible to derive, beyond the causal propertied, the temporal properties of the "obstacle-avoidance" state from the interaction of the (physically) implemented states.

[13] Many parts of our cognitive system have especially developed to meet time constraints of the environment. There is evidence for neural as well as chemical internal clocks (that work at certain clock rates), oscillator circuits that adapt to external cycles, etc. None of this would work if the system ran at 1/10000th of its regular speed. The same is true for digital circuits that have been designed to work at certain clock rates.

distance in time between causally connected states), then this would in theory suffice to capture an essential aspect of possible physical that implementations of the functional architecture. It would, for example, allow us to model the functional architecture computationally, i.e., to implement a virtual machine that abides to the temporal constraints (as many computational descriptions can handle temporal metrics, just take programming languages for real-time systems).

Behavioral states, therefore, seem to be an abstraction, which can be implemented computationally, and thus realized physically on computational systems. At the same time, behavioral states are abstract enough to capture aspects of minds that seem to be intrinsically connected to their causal structure and not to their physical realization ("organizational invariants" as Chalmers, 1997, puts it), thereby connecting them to functional descriptions of cognitive architectures.

5 Conclusion

The level of description of behavioral states is *intermediate* and *intermediary*, because it specifies states that could be realized in many different physical ways (in neural architectures, but possibly also in digital ones, and others), yet retains at least one crucial physical and causal aspect not retained in functional states: time! By explicitly incorporating time and thus allowing for modeling the temporally extended interactions between different states, this level might not only prove useful for constructing systems that exhibit complex causal interactions (such as minds), but also for explaining *how* functional states are related to physical states by viewing them as (not necessarily disjoint) collections of behavioral states.

References

Arkin, R. C. Motor Schema-Based Mobile Robot Navigation. *International Journal of Robotic Research*, Vol. 8, No. 4: 92-112, 1989

Block, N. What is Functionalism? *The Encyclopedia of Philosophy Supplement*, Macmillan, 1996

Brooks, R. A Robust Layered Control System for a Mobile Robot. *IEEE Journal of Robotics and Automation*, Vol. RA-2, 1: 14-23, 1986

Chalmers, D. J. A computational Foundation for the Study of Cognition, 1997 (published on the internet)

Kim, J. *Philosophy of mind*, Westview, 1996

Lorenz, K. *The foundations of ethology*, New York, Springer Verlag, 1981

Lorenz, K. and Leyhausen, P. *Motivation and Animal Behavior: An Ethological View*, Van Nostrand Co., New York, 1973

McFarland, D. *The Oxford Companion to Animal Behavior*, Oxford University Press, 1981

Port, R. and van Gelder, T. *Mind as Motion: Explorations in the Dynamics of Cognition*, MIT Press, Cambridge, 1995

McClelland, J. L. and Rumelhart, D. E. *Parallel Distributed Processing*, Vol. 1 and 2, MIT Press, Cambridge, 1986

Pfeiffer, R. and Scheier, Ch. *Understanding Intelligence*, MIT Press, Cambridge, 1999

Scheutz, M. Implementing Functional Architectures? 2000a (Submitted to VI. Congress of the Philosophical Society)

Scheutz, M. Surviving in a Hostile Multi-Agent Environment: How Simple Affective States Can Aid in the Competition for Resources, *Proceedings of the Thirteenth Canadian Conference on Artificial Intelligence*, Springer Verlag, 2000b

Sloman, A. Supervenience and Implementation: Virtual and Physical Machines, 1998 (submitted to ECAI)

DRAFT
Future Models for Mind-Machines

Marvin L Minsky
Massachusetts Institute of Technology
http://www.media.mit.edu/people/minsky/

Abstract

"Seek not to follow in the footsteps of men of old; seek what they sought." – **Matsuo Basho**

In recent years some political leaders of several countries have expressed concern that in future years their countries will not have enough young people to support the large proportion of old ones. So, incredibly, they propose to take steps to increase their reproduction rates. Instead we embark on programs to develop intelligent robots that could increase productivity in the fields where shortages may appear. Then, each working human could easily support many more other ones – with less damage to our environment. However, there has not been much progress in recent years toward making machines that are able to do most mundane jobs that people do. I think this is because most AI researchers have not used adequate large-scale models for designing systems that could have enough "common sense" or "resourcefulness."

1 Introduction

Humanity has always faced new technological frontiers – but rarely did it appreciate those wonderful opportunities. However, the past three centuries has been different, I think – and over the past fifty years, we've seen the most immense progress in history. For example Physics, Astronomy, and Cosmology have progressed perhaps more in the past half-century than they did since Galileo's time. Biology has moved even more quickly; the field of molecular Biology was virtually born just fifty years ago. Today, I think, we are entering a similar phase of Psychology.

To build reliable, humanlike robots, we'll need ways to make them understand the problems that we want them to solve. One way to do this would to enable them to think in ways like ours. However, we don't yet know how to do this – because we still know too little about our own minds. Our minds are working all the time, but we rarely think about what minds are. What are minds made of and how they work? How do minds build new ideas? Why could our scientists discover so much about atoms and oceans and planets and stars – yet so little about what our feelings are? Our minds are working all the time – yet we know almost nothing about them. We rarely discuss these subjects in schools, or think about them in our daily lives. It is almost as though we've imposed a taboo against trying to think about such things.

How does Imagination work? How do minds learn from experience? How do we recognize what we see? How do we choose which words to say? How do we understand what they mean? How does commonsense reasoning work? Each of these common abilities is based on huge networks of processes. So, to answer those questions, we'll need to accumulate more good ideas about what *are* those networks, how they evolved, and how their resources have managed to merge – to form the constructions we call our minds. In this essay I will start by reviewing some ideas about minds – each of which has just enough parts to answer certain kinds of questions. Then I will suggest how these simple models can be expanded and combined to make better theories about our psychology. (Each brief section below will be further discussed in my forthcoming book, *The Emotion Machine*.)

2 One-Part Models of Mind

The most popular concept of a human mind envisions each person as having a 'Self" – which embodies all those features and traits that distinguish you from everyone else. But when we ask what Selves actually do, we're likely to hear this vacuous view:

Your Self views the world by using your senses, and chooses all your desires and goals. Then it solves all your problems for you, by exploiting your 'intelligence'. It formulates plans for what next you should do – and then makes the pertinent muscles contract so that your body performs your acts.

Isn't this a strange idea? It says that you make no decisions yourself but just delegate them to something else – to that mythical person you call 'your Self'? Clearly this 'theory' can't answer our questions – so why would our minds concoct such a fiction?

Therapist: "That simplistic legend makes life seem more pleasant. It keeps us from seeing how much of our soul is controlled by unconscious, detestable *goals."*

Pragmatist: "It also helps to make us efficient! More complex ideas might just slow us down. It would take too long for our hardworking minds to understood everything all the time."

The trouble with that "Self" idea is that it does not explain what's inside a mind. It's a theory that doesn't have enough parts we can use to build explanations. If you ask about how your mind makes decisions, the Central-Self model just avoids that question, by ascribing all your abilities to another mind inside your mind. (Before the dawn of modern genetics, a similar theory was prevalent: it proclaimed that every sperm already contained a perfectly formed little personage.) The notion of a Central Self can't help us to understand ourselves.

Many other popular theories try to derive all the virtues of minds from one single source or principle:

Survival instinct: All our goals stem from the instinct to survive.

Pleasure Principle: All our drives are based on seeking pleasure

Aversion Principle: We're driven by needs to escape from pain.

Conflict Resolution: All our actions are directed at resolving conflicts.

Urge to control: Our resources evolved to control our environment.

Reinforcement and Association: The mind grows by accumulating various kinds of correlations.

Each of these 'unified theories of mind' has virtues and deficiencies. For example, the Survival-Instinct hypothesis helps to describe a wide range of behaviors – but it's based on a wonderfully wrong idea. Over the course of our evolution, our brains assembled a great host of systems – each of which served in a separate way to protect us from certain kinds of harm. The result of the process was that a brain is a 'suitcase' of systems with similar functions; however, those systems have no common structure – so to understand how those systems work, we'd have to examine them one by one. That 'survival instinct' is just an illusion. When you look at mind as a single thing – instead of a grand architectural scheme – you'll see little more than a featureless blur, instead of the marvelous structure you are.

3 Two-Part "Dumb-Bell" Models of Mind

Many popular mental models are based on "dumb-bell" distinctions that try to divide the entire mind into just two complementary portions, such as *Left-Brain vs. Right-Brain, Rational vs. Intuitive; Intellectual vs. Emotional, or Conscious vs. Unconscious.* These can be better than Single-Self models. However, they too often support old superstitions that make it hard to develop more useful ideas. For example, when neurologists discovered some differences between the brain's two hemispheres, this revived many views of our minds which were, in our more ancient times, expressed in terms of opposites like Devils vs. Angels, Sinners vs. Saints, and *Yins* vs. *Yangs*. So this pseudoscientific scheme revived nearly every dead idea of how to see the mental world as a battleground for two equal and opposite powers.

Why are dumbbell theories so popular? I suspect that this is because – just like those old myths – they provide just enough parts to tell stories of conflicts. Instead of believing such story-like myths, we should try to make theories of why they enchant us.

4 Three-part Models of Mind

Three-part theories, although still too simplistic, are rich enough to suggest better ideas. Here are a few of my favorite such frameworks:

Paul MacLean's "Triune brain" hypothesis [*The Triune Brain in Evolution*] tries to explain how minds behave in terms of machinery that evolved in three stages – namely, when our ancestors became Reptiles, then Mammal, and finally, Primates. He identifies those hypothetical 'layers' with stages of our evolutionary history – as well as with different aspects of thinking. However the evolution of our 'lower' brain systems did not suddenly cease when those 'higher' ones came. They all continued to co-evolve, so that each of our behavioral functions is based on components from every stage.

Eric Berne's "Transactional Analysis" hypothesis is based on the idea that every person evolves subpersonalities based on models of the child, adult, and parent. [Eric Berne, *Transactional Analysis in Psychotherapy*] This is quite different from MacLean's scheme, and more suitable for describing the development of social behaviors.

Sigmund Freud's "Psychoanalysis" theory was based on a psychological triad of interactions between a "Id" or collection of Instinctive urges, a "Superego" that embodies our high-level socialized goals and prohibitions, and an "Ego" that resolves or suppresses the conflicts between them. I especially like Freud's 'sandwich-like' architecture, first because it is non-hierarchical, and second because it emphasizes 'negative knowledge' – that is, knowing which things one should not do. Competence requires both positive and negative knowledge – and I suspect that as much as half of our commonsense knowledge may have of this negative character. [See Marvin Minsky, "Negative Expertise"]

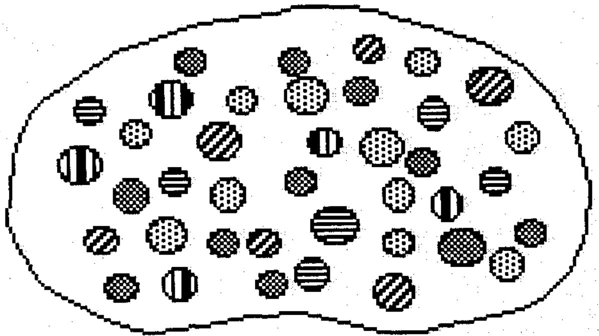


Figure 1

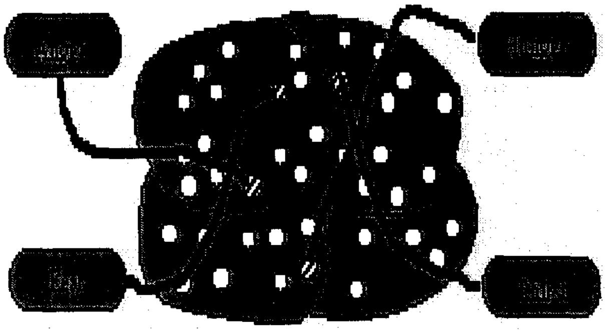


Figure 2

5 Viewing the Mind as a "Cloud of Resources"

The human brain has hundreds of parts that have different functions – so any comprehensive model of mind must include descriptions of all those resources. By "resources" I mean to include both bodies of knowledge and program-like processes – such as perceptual schemes for making descriptions, for forming goals and for making decisions, or methods for solving difficult problems. Especially, the brain needs resources to assess what other resources do – e.g., to decide which ones are making good progress or wasting our time, or to recognize conflicts and try to resolve them. This suggests that we think of the brain as a cloud of varied resources, where each can use others in certain ways. [Figure 1]

Holistic Philosopher: That whole idea seems wrong to me. By dividing the mind into smaller parts, aren't you likely to miss the whole point? Unless you look at a thing as a whole, you'll miss its most vital aspects. Surely you need a more holistic view.

Every representation we use is bound to miss some important aspects, for which we must switch to another view or a different type of representation. So to understand anything well, we'll usually need to use several such views, and some ways to interconnect them. Certainly, this must include some "high level" views that try to describe the entire thing. However, 'holistic thinkers' don't always recognize that vague summaries have their limits, too. Like cartoons, they give us illusions of "seeing the whole thing at once." However, these tend to be oversimplified views that cannot explain anything in detail – just as maps display only a few striking features, while suppressing details of the actual regions.

This idea of a mind as a cloud of resources might seem too vague to have much use, but it helps us to escape from those dumb two-part models. For consider the following type of phenomenon: One moment your baby seems perfectly well, but then come some restless motions of limbs. Next you see a few catches of breath – and in just a few moments the air fills with screams. The Single-Self model has no way to explain what could possibly cause such changes – but this is easier to explain if we assume that an animal's brain contains several almost-separate sets of resources – where each set evolved to serve some vital need like procreation, nutrition, or defense. This model, developed by Nikolaas Tinbergen and Konrad Lorenz, is described in Tinbergen 's book "The Study of Instinct". It does not explain much about human thought but has turned out to be surprisingly good at accounting for much of what animals do.

One form of a system with such a description might resemble a human community, where different people do different jobs – as in Howard Gardner's theories about Multiple Intelligences, which lead to good models for representing a person's largest scale behavior. [See, for example, Howard Gardner, *Frames of Mind: The Theory of Multiple Intelligences.*] However, each member of a human family, village, or corporation is already a competent and autonomous person – whereas inside a single person's brain, each resource is far more specialized; it can do only a certain few things, and depends on the rest for everything else. So, when we envision an individual human mind, it may be better to think of a large network of smaller machines. Of course the resource-cloud view is not quite what one would call 'a theory' – because while it says that the system has parts, it does not specify what those parts are. It says they're connected, but doesn't say how. It suggests no particular architecture. However, the very vagueness of the Resource-Cloud idea is what makes it a powerful tool for thought, just because it reminds us of those deficiencies.

In particular, it suggests that the brain must contain enough "managers" to monitor, supervise, appraise, and control the activities in particular sets of other resources. A typical resource is connected to several others and can use those connections in various ways, e.g., to exchange some information with them, to exploit them for various purposes. In particular, some resources will be especially equipped to turn some other resources on or off. Thus, from every moment to the next, only certain resources will be active – and these will determine what your mind does at any particular moment of time. This suggests a theory of emotions in which each emotion or 'disposition' results from some more or less persistent arrangement in which certain resources are highly active, while others are more

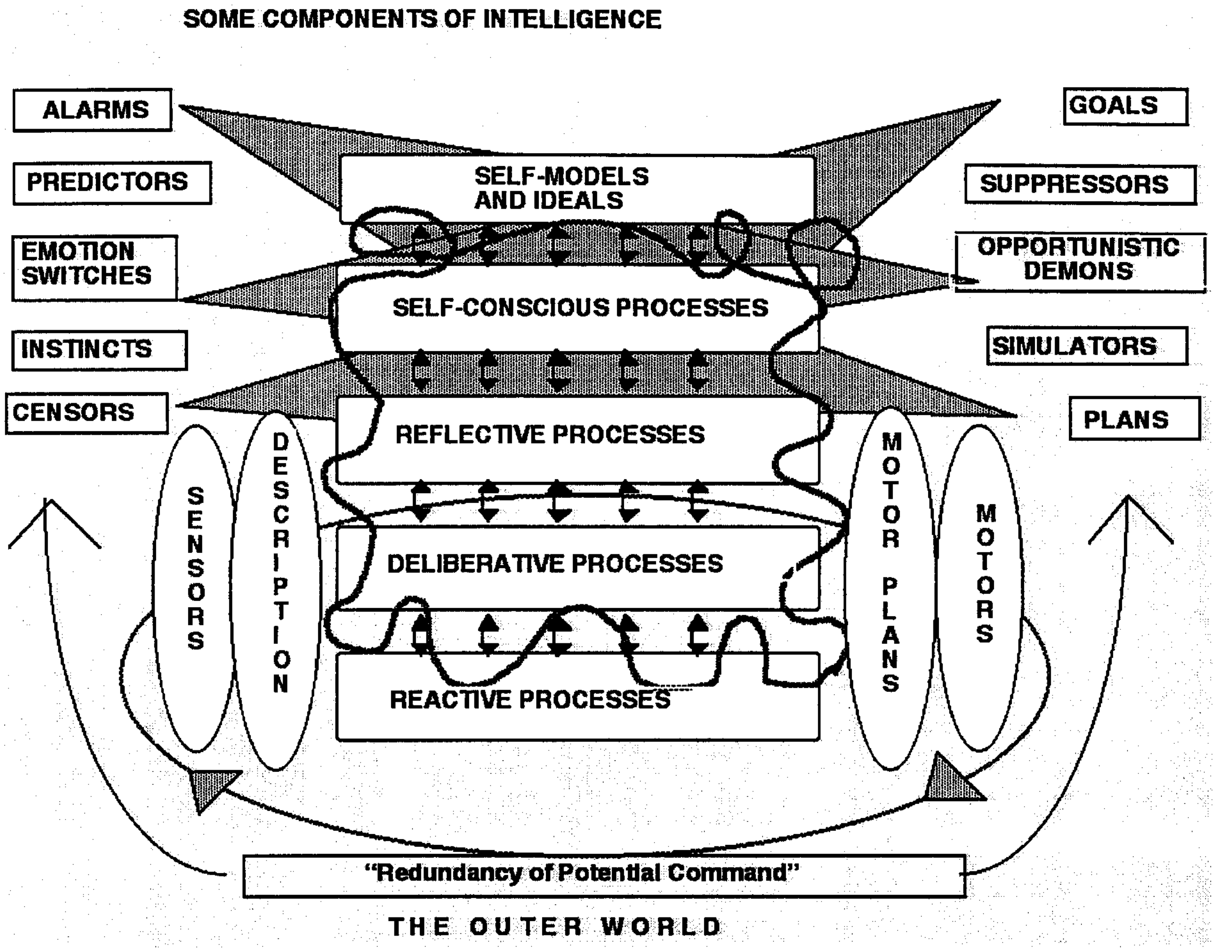


Figure 3

quiescent:

An emotional state is what happens when we 'turn on' a certain large set of resources. [Figure 2]

6 A Large-Scale Model of Conscious Thought

How does a brain employ its resources? One way to start would be to assume, as has been suggested by Aaron Sloman, that our resources are arranged in three or more levels [Figure 3]:

– A "reactive" collection of resources "A" that includes systems for memory, perception, and other procedures, etc.

– A "deliberative" collection of resources "B" that observe and react to the activities in A.

– A "self -reflective collection of resources that observe and react to what happens in "B," etc.

No such a complex system could work without more machinery to control it. To see what that management might involve, let's look at one fragment of human behavior.

Joan is part way across the street on the way to present her finished report, and she's thinking about what to say at the meeting. She hears a sound and turns her head to see a quickly oncoming car. Uncertain whether to cross or to retreat, but uneasy about arriving late, she elects to sprint across the road. Later she reflects about her rather reckless decision. "I could have been killed if I'd missed my step – and then what would my friends have thought of me?"

Every minute of every day, we experience streams of events like these. To some of them, we react without thinking. To others we act more deliberately. Let's try to imagine what goes on in Joan's mind as she makes her way to that meeting.

Reactive Awareness: *She hears a sound and turns her head in that direction.*

When Joan turned her head to look around, was she conscious of that sound, or was that a 'mindless' reaction? Was she aware of which muscles she used to make herself walk across that road? Not likely, because most of us don't even know which muscles we own. Other resources inside Joan's brain must be more involved with such affairs – but because no pathways communicate this, Joan is not 'aware' of this. What is awareness, anyway? What determines its focus and range? What machinery does it use in the brain? How many things can you do at one time – and how many can you be aware of? Presumably, that will depend on the extent to which they each use different resources for them. But when Joan perceives that approaching car, this quickly takes the center stage and takes hold of her full attention.

Deliberative thinking: *She is thinking about what to say at the meeting.*

To do this she must first consider how several alternatives might be received – and then compare those imagined reactions. This may require so many resources that she has to do this sequentially.

Reflective reasoning: *Joan reflects about what she has done, and concludes that she made a poor decision.*

To what extent was she aware of what determined her risky decision? Reflection involves thinking about what one's brain's has recently done. That kind of reflection requires resources to examine the records that other resources have been keeping.

Internal "Meta-Management": *but uneasy about arriving late*

Another family of resources is monitoring Joan's temporal progress, and decides that whatever the merits of what she is thinking, she cannot afford to delay her decision.

Self-conscious Reflection: *"What would my friends have thought of me?"*

Joan thinks about how her friends might change their mental representations of her. Reflections like this have as their subject, a person's private self-representations – the models or self-images that we all construct to describe ourselves.

So the architecture of our minds must include at least these five kinds of layers. This idea is further developed in my forthcoming book *The Emotion Machine.* Of course, a real brain is far more complex, and each of those layers and arrows eventually must be replaced by hundreds of smaller components, interconnected by thousands of pathways. (This scheme is partly inspired by the research of Aaron Sloman.)

7 Psychology Needs a Network of Large Scale Models

To understand the human mind, we'll need to use several kinds of models. Some will need only a few parts – enough to answer just certain questions – but others will have to be much more complex, to explain such 'higher mental functions' as reasoning, imagination, decision-making, and consciousness. And, since no one such vision can explain everything that we want to explain, we'll have to keep switching between different models.

Critic: That sounds very disorderly. Why can't you simply combine them all, like the physicists try to do, into a single one that combines the virtues of all those theories?

That would result in such a mess that no one could hold it in mind all at once. We have to be able to use different views to highlight different aspects of things, and that's why we still tend to speak about Physics, Chemistry, and Biology – as though these were more or less separate subjects. Some of the contents of each of those fields can be deduced 'in principle' from more basic physical principles. The trouble is that we can't do this "in practice" because no one can actually solve those equations. (And in Psychology, we can't expect to have any such set of equations.)

The 'large-scale models' that we've described are not 'hypotheses' to prove false or true. Instead, they are more like 'points of view' – particular ways to think about things, or to focus attention on various problems. So it's not a question of which one is 'right', but where and when to use each view. Each is a rough architectural plan that will help us to understand certain things. However, because each of them has limitations, we'll have to keep changing our points of view, by shifting between different Large-Scale Model. Our own human brains are too complex for us to envision all at once – so we'll have to keep changing our representations. This shifting around might at first seem disturbing, but later we'll see that it's worthwhile – because it will also enable us to describe the process that actually happens *inside* our minds!

Using multiple models is not just a way to state theories about psychology. It is part of psychology itself – because we can only understand complex things by switching between different representations. This is the basis of our most powerful way to think: to keep interweaving different views so fluently that we never suspect that we're doing it.

No system as complex as a human mind can be well described by a few simple rules – because each rule would have many exceptions. This is because each part of such a system is likely to reflect the particular ways that it once worked in the environment in which it evolved (both out in the world and inside the brain). Then whenever some subsystem fails to work, those brains will tend to evolve a 'patch' – an 'ad hoc' way to help it to work. The result is the accumulation of multiple layers of patches, over hundreds of megayears of evolution.

What does it mean when you say to yourself, "That was a stupid thing to do," or "I didn't expect to succeed at that!" You're always praising or blaming yourself, and holding yourself responsible. But whenever you change your emotional states, you're using some different processes and memories – so you are no longer the very same 'you'. What gives us the sense that we remain the same while shuttling among those states? Partly this must be because we use the terms for describing ourselves. Terms like 'me', 'myself' and 'I' help us to envision ourselves as like the 'eye' of a cyclone that stays in one place while everything circles around it. In *The Emotion Machine* I'll argue that the mind has no single well-defined thing that remains the same while controlling the rest. Instead we each have a rich collection of personal, large-scale models of ourselves.

Our 'commonsense' ideas about ourselves have so many bad misconceptions. We all have grown up with certain traditions that tacitly assume, for example, that we each 'hold' a single set of beliefs. Thus, when someone asks what you "really" believe – or what your 'true' intentions are – or what you 'really' meant to say – those phrases make sense in the Single-Self realm. But a realistic view of your mind would show how it uses at various times, different arrangements of its resources – each of which can make you exhibit different opinions, ideas, and convictions. And despite what each of us likes to think, no particular one of those cliques deserves to be called "what I truly believe".

8 Advice to Students

How should student select a career in these future burgeoning technical fields? One approach is to ask what is the most popular field now. Another approach is the opposite: to choose an underpopulated area. Now, the popular fields offer great current opportunities. (For example, in genetics, each of our hundred thousand genes may take a few lifetimes to understand – for evolution has used all the tricks that the physical world permits.)

However, a young, ambitious student who wishes to make a great and fundamental contribution should consider the idea of deliberately avoiding the most popular fields! For, consider the arithmetic. Imagine that in the next ten years there will be ten major discoveries in a certain field where already ten thousand researchers are working. (This is the case at present in such areas, for example, as Neural Networks, Genetic Programming, Simple Mechanical Robots, Statistical Linguistics, and Statistical Information Retrieval.) Then in each decade, each of those researchers will have perhaps one chance in 1,000 to make a major discovery. Contrast this with the situation in an equally important field that currently employs only the order of a dozen good researchers – as in the areas of *Representing Commonsense Knowledge* or *Large-Scale Cognitive Architectures*. Then you'll have a thousand times better chance to make an important discovery! Many students have complained to me that it's easier to get a job in a currently popular field. However, if one looks for less faddish alternatives, one may find that the competition is accordingly less.

References

Eric Berne, *Transactional Analysis in Psychotherapy*, Grove Press, New York, 1961

Howard Gardner, *Frames of Mind: The Theory of Multiple Intelligences*, Basic Books, 1993, ISBN: 0465025102.

Paul MacLean *The Triune Brain in Evolution*, Plenum Pub Corp., ISBN: 0306431688, New York, 1990

Marvin Minsky, Negative Expertise, *Int. J. Expert Systems*, 7,1, pp. 13–19, 1994 or www.media.mit.edu/people/minsky/papers/NegExp.mss.txt

Nikolaas Tinbergen The Study of Instinct, Oxford University Press, 1951

AISB'00 Convention Invited Talk

From intelligent organisms to intelligent social systems: how evolution of meta-management supports social/cultural advances.

Aaron Sloman
School of Computer Science, The University of Birmingham
http://www.cs.bham.ac.uk/~axs/

Abstract

This invited talk will speculate on ways in which a type of three level information processing architecture including reactive, deliberative and meta-management layers, can support and be influenced by social interaction.

1 Introduction

It is now fairly common in AI to think of humans and other animals, and also many intelligent robots and software agents, as having an information processing architecture which includes different layers which operate in parallel, and which, in the case of mammals, evolved at different stages.

The idea is also quite old in neuroscience. E.g. Albus (1981) presents MacLean's notion of a layered brain with a reptilian lowest level and at least two more recently evolved (mammalian) levels above that. AI researchers have been exploring a number of variants, of varying sophistication and plausibility, and varying kinds of control relations between layers.

In our own work (e.g. Sloman (2000)) we have assumed a coarse threefold sub-division between concurrently active reactive, deliberative and meta-management (reflective) layers, all operating partly independently of the others, all with specialised sensory inputs from layered perceptual mechanisms, and with access to a hierarchical motor control system. Different classes of mental processes, including motivations, moods, emotions and types of awareness depend on the different layers.

The meta-management layer, which evolved latest and is rarest in animals, is assumed to be able to monitor, categorise, evaluate, and to some extent control other layers, e.g. redirecting attention or altering the mode of deliberation, though it may sometimes be disrupted by other mechanisms, e.g. in emotional states where attention is repeatedly drawn to an object or topic of concern, even against one's will.

2 Executive function

The common reference to "executive function" by psychologists and brain scientists seems to conflate aspects of the deliberative layer and aspects of the meta-management layer. That they are different is shown by the existence of AI systems with sophisticated planning and problem solving and plan execution capabilities without meta-management (reflective) capabilities.

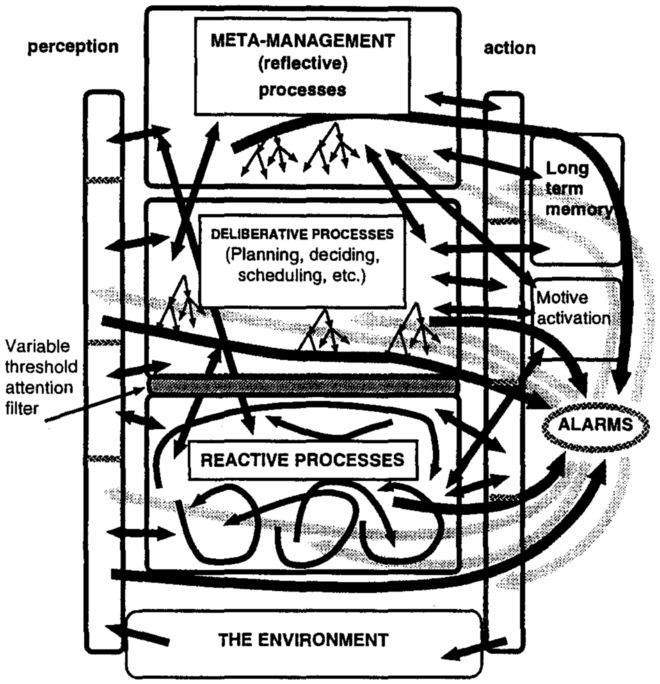
A symptom would be a planner that doesn't notice an obvious type of redundancy in the plan it produces, or which can solve many problems but does not learn by noticing patterns in its own performance, so that it repeats old mistakes, and doesn't recognise cases where it is going round in circles trying to solve a hard problem.

One consequence of having the third layer is the ability to attend to and reflect on one's own mental states, which could cause intelligent robots to discover qualia, and wonder whether humans have them.

3 Changing personalities

There is some evidence that in humans the third layer is not a fixed system: not only does it develop from very limited capabilities in infancy, but even in a normal adult it is as if there are different personalities "in charge" at different times and in different contexts (e.g. at home with the family, driving a car, in the office, at the pub with mates).

Taking most of that for granted, this talk will speculate about the influence of a society or culture on the contents and capabilities of the third layer in humans. The existence of such a layer does not presuppose the existence of an external human language (e.g. chimpanzees may have some reflective capabilities), though it does presuppose the availability of some internal formalism, as do the


perception
META-MANAGEMENT (reflective) processes
action
Long term memory
DELIBERATIVE PROCESSES (Planning, deciding, scheduling, etc.)
Motive activation
Variable threshold attention filter
ALARMS
REACTIVE PROCESSES
THE ENVIRONMENT

The Birmingham Cogaff Architecture

This architecture is described in more detail in the Introduction to the Symposium on How to Design a Functioning Mind at this Convention. It is assumed to have a reactive layer, a deliberative layer capable of doing 'what if' reasoning and a meta-management layer capable of monitoring, categorising, evaluating, and to some extent controlling and redirecting processes in other parts of the system (though not all). To make all this work many additional components are required, not shown here.

reactive and deliberative layers.

When an external language develops, *one* of its functions may be to provide the categories and values to be used by individuals in judging their own mental processes (e.g. as selfish, or sinful, or clever, etc.)

This would be a powerful form of social control, far more powerful than mechanisms for behavioural imitation,. for instance. It might have evolved precisely because it allows what has been learnt by a culture to be transmitted to later generations far more rapidly than if a genome had to be modified. However, even without this social role the third layer would be useful to individuals, and that might have been a requirement for its original emergence in evolution.

If true this could have important implications for AI researchers working on multi agent systems, as well as philosophers, brain scientists, psychiatrists, social scientists and biologists studying evolution.

4 Mechanisms required

It is conjectured that there is a collection of different, culturally influenced, 'personae' which take control of the top layer at different times, e.g. when a person is at home with family, when driving a car, when interacting with subordinates in the office, in the pub with friends, etc.

For such a thing to be possible, it seems that the architecture will require (a) something like a store of 'personalities', (b) mechanisms for acquiring new ones or modifying and extending old ones (e.g. via various social processes), and (c) mechanisms for retrieving and activating personalities when relevant, e.g. allowing changes in external contexts to 'switch control' between personalities.

5 Disorders of meta-management

There are many ways in which such a system can go wrong, or break down. In particular some forms of pathological disorders may be connected with malfunctions in the processes of switching between different personalities. Others may be concerned with the process of acquiring and shaping personalities. There could also be ways in which stored personalities get corrupted.

In less extreme cases the manifestation may be an inability to get on well with other people, escalating aggression or depression resulting from the wrong sort of personality being invoked, etc. In such cases "milder" treatments such as counselling or behaviour therapy may help.

6 Cultural influences

The influences of a culture on an individual are diverse and include determining which ontology the individual develops for categorising both things in the environment and its own internal states, providing notations and languages for expressing the ontology and reasoning about entities within it, influencing collections of beliefs about those ontology, producing collections of external behaviours and thinking strategies, and influencing the standards and values deployed in generating new goals and choosing between competing goals.

One of the questions that arise for engineers interested in producing intelligent systems is whether these processes will occur in intelligent artefacts. Insofar as programming every detail of a sophisticated robot or software agent may be impossibly complex it may be essential that a bootstrapping mechanism be provided by which it can learn for itself.

However the need to be able to cope with unforeseen situations and unexpected conflicts will require the robot not only to learn standard strategies and behaviours, but also values, standards, and modes of thinking about and dealing with conflicts of values.

It could turn out that the most effective way of enabling them to do this well is to mimic the human architecture. In that case we may also expect some of them to develop human forms of fallibility and pathologies, including both those which arise out of physical damage to the underlying mechanisms and those which arise out of 'software bugs' due to a history of bad experiences.

References

James S. Albus. *Brains, Behaviour and Robotics.* Byte Books, McGraw Hill, Peterborough, N.H., 1981.

Aaron Sloman. Architectural requirements for human-like agents both natural and artificial. (what sorts of machines can love?). In Kerstin Dautenhahn, editor, *Human Cognition And Social Agent Technology*, Advances in Consciousness Research, pages 163–195. John Benjamins, Amsterdam, 2000.

Inspiration from Neurosciences to emulate Cognitive Tasks at different Levels of Time

Frédéric ALEXANDRE
LORIA-INRIA; BP 239; F-54506 Vandoeuvre Cedex; falex@loria.fr

Abstract

Our team has been working for more than ten years on the modelling of biologically inspired artificial neural networks. Today, our models are used to different cognitive tasks like autonomous behavior and exploration for a robot, planning, reasoning, and other tasks linked to memory and internal representation building. We present the framework that underlies these models through the time delays related to several fundamental properties like information coding, learning, planning, motivation.

1 Introduction

The goal of this poster is to present one original consequence of getting inspired by biology to elaborate temporal mechanisms for behavioral modelling. More precisely, symbolic or numerical tools for temporal processing are generally applied to very specific tasks like for example high level planning, perceptive scene analysis or temporal alignment of speech signal. It is then quite impossible to integrate these models to get the corresponding full range of properties, necessary for the implementation of a realistic task including these various levels of time. One important advantage of using biologically inspired neural networks for temporal processing is that biology offers a complete framework of inspiration from the lowest to the highest levels of time scale.

2 The bit level

At the level of 10^0 millisecond, a neuron can perform synaptic transmission to its closest neighbors. One millisecond is also the duration of a spike. At the neuronal level, this time scale is thus the level of the lowest bit of information. This level of neuronal processing is deeply studied in the so-called "spiking neuron" approach Maass and Bishop (1998). Here, a model of neuron is considered as an elementary unit emitting spike trains. At the synaptic level, Grossberg (1984) proposes differential equations to model neurotransmitter consumption and production, whose equilibrium state yields a non-linear function, similar to classical sigmoidal transfer function.

3 The coding level

At the level of 10^1 milliseconds, the duration of the interval between two spikes can be evaluated, since, in the central structures, the neuronal maximal frequency cannot exceed 100 Hz. Whereas classical connectionist approaches Hertz et al. (1991) generally use continuous models of neurons, whose activation value corresponds to the estimated mean frequency of spike trains, the spiking neuron approach deepens the idea that, as spike emission is a binary process, all the neuronal information is included in the timing of the spikes. Beyond simple frequency estimation, other rate coding or even phase coding can be also investigated at this level of description Maass and Bishop (1998). At the behavioral level, this time scale corresponds to inter-areal communication including for example feedback information and focus of attention.

4 The processing level

At the level of 10^2 milliseconds, the activation dynamics of a population of neurons can stabilize into a synchronized state. This phenomenon can be precisely studied with spiking neurons. For example, Mar et al. (1999) investigates how a population of coupled model neurons can perform noise shaping. The population of neurons is such an important and consistent level for neuronal processing that several researchers like Edelman (1987); Burnod (1989); Alexandre et al. (1991) have chosen this level of description to define an integrated neuronal automaton which thus corresponds to a synchronized population of neurons.

At the behavioral level, this time scale corresponds to the stabilization of activity consecutive to oscillations created by sensory and motor events, from the first to the last processing layer. This duration can thus allow to perform recognition or action in elementary sensorimotor loops, thanks to the integration of activity in these layers. Koechlin and Burnod (1996) describes this phenomenon in models ranging from the spiking neuron to the integrated automaton levels.

5 The learning level

At the level of 10^3 milliseconds, neurons in the highest levels in the associative cortex can stay active for such a duration. This lasting internal representation can allow for such process as object exploration, including multimodal dimensions. Kosslyn et al. (1992) describes how a pattern can be recognized through the identification of its subparts (temporal areas) together with their localization (parietal areas).

This time scale also corresponds to learning elementary processes. From the basic idea proposed by Hebb Hebb (1949) as soon as 1949, stating that reinforcement can be produced by presynaptic and postsynaptic activity coincidence, many elaborated learning rules have been proposed. Among them, some try to integrate a temporal dimension to this rule, allowing presynaptic and postsynaptic activities to be consecutive and not simultaneous. These rules generally use the trace signal principle Reiss and Taylor (1991), yielding a lasting and progressively extinguishing activity when the signal is no longer present. This lasting activity can make two separate signals meet and perform learning. This idea was for example exploited in Sutton and Barto (1981) to model pavlovian conditining.

6 The stack level

At the level of 10^4 milliseconds, neurons in the frontal cortex can have a sustained activity which is the basis for working memory in this region. Burnod (1989); Fuster (1996) describe how the control of bistable activity in frontal neurons can allow to build stacks that can command the triggering of sensorimotor events in the posterior part of the cortex. More precisely, the temporal organization of behavior can be performed with such a mechanism, as shown by computer science implementation by Guigon et al. (1995) for monkey conditioning paradigm modelling or by Frezza-Buet and F. (1998) for environment exploration by an autonomous robot.

7 The modulation level

At the level of 10^5 milliseconds and more, rhythms can be produced by extra-cortical structures like the reticular formation or the hypothalamus. Neuronal intrinsic metabolic and genetic processes can occur at such very long time constants and influence cortical activity. This level can thus be defined as the level of emotion, motivation, mood and other global influences that can regulate the whole behavior. As proposed in Burnod (1989), such modulatory phenomena can be modelled through global variables, able to influence the whole network.

References

F. Alexandre, F. Guyot, J.-P. Haton, and Y. Burnod. The Cortical Column: A New Processing Unit for Multilayered Networks. *Neural Networks*, 4:15–25, 1991.

Y. Burnod. *An adaptive neural network: the cerebral cortex.* Masson, 1989.

G. Edelman. *Neural Darwinism: the theory of neural group selection.* Basic Books, 1987.

H. Frezza-Buet and Alexandre F. Selection of action with a cortically-inspired model. In *Seventh European Workshop on Learning Robots*, pages 13–21, 1998.

J. M. Fuster. Frontal lobe and the cognitive foundation of behavioral action. In A.R. Damasio, H. Damasio, and Y. Christen, editors, *Neurobiology of Decision-Making*. Springer, 1996.

Stephen Grossberg. Some normal and abnormal behavioral syndromes due to transmitter gating of opponent processes. *Biological Psychiatry*, 19(7):1075–1117, 1984.

E. Guigon, B. Dorizzi, Y. Burnod, and W. Schultz. Neural correlates of learning in the prefrontal cortex of the monkey: A predictive model. *Cerebral Cortex*, 5(2): 135–147, 1995.

D. O. Hebb. *The organization of behaviour.* Wiley, New-York, 1949.

J. Hertz, A. Krogh, and R. Palmer. Introduction to the theory of neural computation. Addison Wesley, 1991.

E. Koechlin and Y. Burnod. Dual population coding in the neocortex: A model of interaction between representation and attention in the visual cortex. *Journal of Cognitive Neurosciences*, 8:353–370, 1996.

S. Kosslyn, C. Chabris, C. Marsolek, and O. Koenig. Categorical versus coordinate spatial relations: computational analysis and computer simulations. *Journal of Experimental Psychology: Human Perception and Performance*, 18:562–577, 1992.

W. Maass and C. Bishop, editors. *Pulsed Neural Networks*. Bradford Book, MIT Press, 1998.

D. Mar, C. Chow, W. Gerstner, R. Adams, and J. Collins. Noise shaping in populations of coupled model neurons. *Proc. Natl. Acad. Sci. USA*, 96:10450–10455, 1999.

M. Reiss and J.G Taylor. Storing temporal sequences. *Neural Networks*, 4:773–787, 1991.

R. S. Sutton and A. G. Barto. Toward a modern theory of adaptative network : Expectation and prediction. *Psychological Review*, 88(2):135–170, 1981.

A Concern-Centric Society-Of-Mind Approach To Mind Design

Steve Allen[1]

Multi-Agent Systems Group,
German Research Centre for Artificial Intelligence,
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany
Email: Steve.Allen@dfki.de

Abstract

In this poster summary, we argue that mental concern-processing mechanisms are amenable to a society-of-mind approach to mind design. We illustrate our case with an information-level analysis of the emotion process, relating the different classes of emotional state to the different layers of our motivated agent framework. We describe how a society-of-mind design-based implementation strategy allows us to add depth to our agent architecture, and incrementally account for more and more of the phenomena of interest. Finally, we report on the results of recent research into the design of cognitively-inspired emotional agent architectures.

1 Introduction

Concerns are broadly defined as dispositions to desire the occurrence, or non-occurrence, of a given kind of situation [Frijda 86, page 335].

Not all concern processing mechanisms need explicit representational forms or structures (as some are emergent), but they do need a systematic framework within which they can be described and operate. In this summary we will use our motivated agent framework (Figure 1) to briefly elucidate the concern-processing mechanisms inherent in the human emotion process.

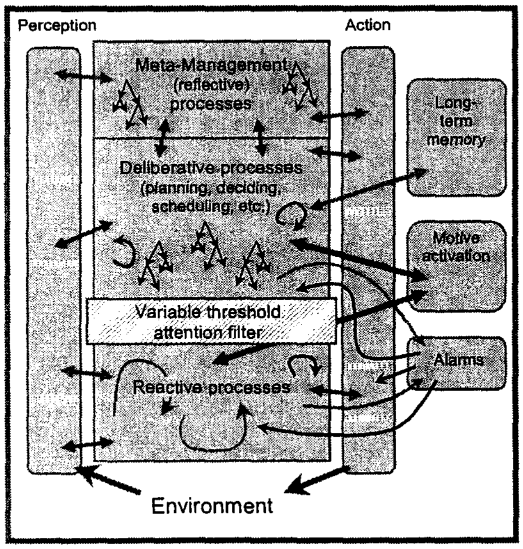

Perception
Action
Meta-Management (reflective) processes
Long-term memory
Deliberative processes (planning, deciding, scheduling, etc.)
Motive activation
Variable threshold attention filter
Alarms
Reactive processes
Environment

Figure 1 Motivated Agent Framework [Sloman 99]

2 Emotional States

By referring different definitions and theories of emotion to the different layers of the motivated agent framework, we can identify three main classes of emotional state [Sloman 99] – *primary*, *secondary*, and *tertiary*.

Primary emotional states: such as being startled, terrified, or sexually stimulated, are typically triggered by patterns in the early sensory input and detected by a global alarm system.

Secondary emotional states: such as being anxious, apprehensive, or relieved, depend on the existence of a deliberative layer in which plans can be created and executed with relevant risks noticed, progress assessed, and success detected. An alarm system capable of detecting features in theses cognitively generated patterns is still able to produce global reactions to significant events in the thought process [see also Damasio 94 and Picard 97].

Tertiary emotional states: such as feeling humiliated, ashamed, or guilty, can be further characterised by a difficulty to focus attention on urgent or important tasks. These emotions cannot occur unless there is a meta-management layer to which the concept of "losing control" becomes relevant.

The three different classes of emotional state should be seen as orthogonal to the common emotion type labels used in everyday language. For example, fear can take the form of a primary, secondary, or tertiary emotion. Each class of emotional state has its own physiological characteristics and hedonistic tone, further underlining the futility of talking about emotional states as active states of a discrete "emotion" system (or systems).

[1] In collaboration with the Cognition and Affect project at Birmingham University.

3 Society-of-Mind

Emotional states are best viewed as an emergent phenomena arising from the interaction of a number of different systems and cognitive processes (only some of which are specific to the generation of emotional states). We can start to make these systems/processes more explicit by mapping their abstract information-processing representations onto our motivated agent framework. This mapping process is performed within the context of Frijda's [86] emotion process, resulting in a generalised design for an emotional agent (see Figure 2).

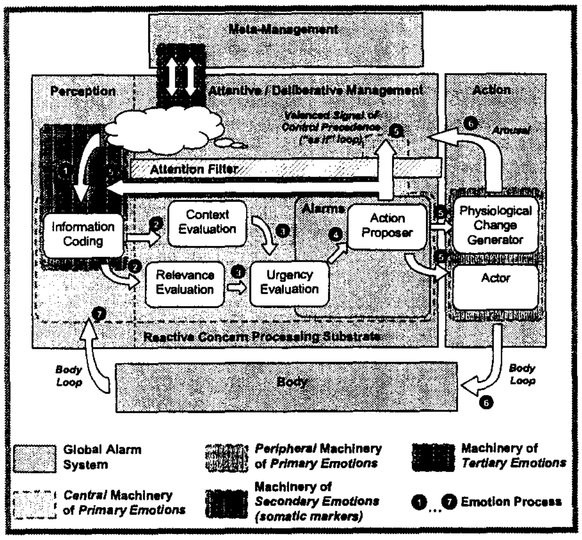

Meta-Management
Perception
Attantive / Deliberative Management
Action
Valenced Signal of Control Precedence ("as if" loop)
5
6
Arousal
1
5
Attention Filter
Information Coding
2
Context Evaluation
3
Alarms
4
Action Proposer
5
Physiological Change Generator
5
Actor
2
Relevance Evaluation
3
Urgency Evaluation
7
Reactive Concern Processing Substrate
Body Loop
Body
Body Loop
6
Global Alarm System
Peripheral Machinery of Primary Emotions
Machinery of Tertiary Emotions
Central Machinery of Primary Emotions
Machinery of Secondary Emotions (somatic markers)
1...7 Emotion Process

Figure 2 – An Information-Level View of the Emotion Process

Having established an abstract design for an emotional agent (noticeably devoid of an "emotion" module), we can now start to refine the architecture through our design-based research methodology – building a series of complete broad-but-shallow implementations of our design to incrementally cover more and more of the phenomena of interest.

We are able to capitalise on the society-of-mind design philosophy by adding depth to our agent designs through the addition of new specialist members within the existing society-of-mind architecture. Furthermore, drawing inspiration from the fields of neurology, we started to map these information-level agents onto regions of the human brain [LeDoux 96, Damasio 94]. For example, Allen [2000] describes the design for an emotional society-of-mind agent architecture, based on earlier work by Cañamero [97] and members of the Cognition and Affect project at Birmingham University [Beaudoin 94 and Wright 97].

4 Conclusions

In this brief poster summary, we have tried to give a flavour of the concern-centric society-of-mind approach we advocate for mind design. Although we have focussed on a single aspect of mind, that of the emotion process, our approach is general enough to apply to other mental phenomena.

Acknowledgements

The author would like to thank Aaron Sloman and members of the Cognition and Affect project at Birmingham University, and the Multi-Agent Systems and Intelligent User Interfaces groups at the German Research Centre for Artificial Intelligence, for their active support of this research.

5 References

Allen, S. (2000). *Concern Processing in Autonomous Agents*. Submitted PhD Thesis. School of Computer Science, University of Birmingham.

Allen, S. (1999). Control States and Motivated Agency. In E. André (Ed.) *Behavior Planning for Life-Like Characters and Avatars: Proceeding of the i3 Spring Days '99 Workshop*. pages 43-61.

Beaudoin, L. (1994). *Goal Processing in Autonomous Agents*. PhD Thesis, School of Computer Science, University of Birmingham.

Cañamero, D. (1997). Modeling Motivations and Emotions as a Basis for Intelligent Behavior. In *Proceedings of the First International Symposium on Autonomous Agents, AA'97*, Marina del Rey, CA, February 5-8, The ACM Press.

Damasio, A. R. (1994, 96). *Descartes' Error: Emotion, Reason and the Human Brain*. London: Papermac. (first published 1994, New York: G. P. Putman's Sons.)

Frijda, N. H. (1986). *The Emotions*. Cambridge: Cambridge University Press.

LeDoux, J. E. (1996). *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*. New York: Simon and Schuster.

Picard, R. W. (1997). *Affective Computing*. Cambridge, MA: The MIT Press.

Sloman, A. (1999). Architectural Requirements for Human-like Agents Both Natural and Artificial. (What sorts of machines can love?). In K. Dautenhahn (Ed.) *Human Cognition And Social Agent Technology*, John Benjamins Publishing.

Wright, I. P. (1997). *Emotional Agents*. PhD Thesis, School of Computer Science, University of Birmingham.

Resource Guided Concurrent Deduction

Christoph Benzmüller*; Mateja Jamnik*; Manfred Kerber*; Volker Sorge†
*School of Computer Science, The University of Birmingham
Edgbaston, Birmingham B15 2TT, England
†Fachbereich Informatik (FB 14), Universität des Saarlandes
D-66041 Saarbrücken, Germany
C.E.Benzmuller|M.Jamnik|M.Kerber@cs.bham.ac.uk; sorge@ags.uni-sb.de

1 Motivation

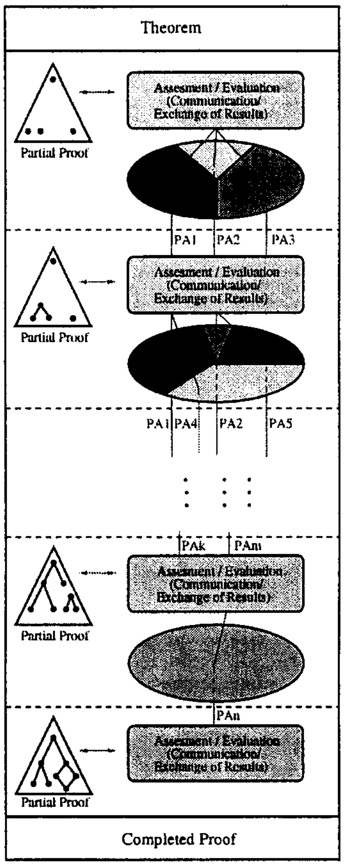
Our poster proposes an architecture for resource guided concurrent mechanised deduction which is motivated by some findings in cognitive science. Our architecture particularly reflects Hadamard's "Psychology of Invention" Hadamard (1944). In his study Hadamard describes the predominant rôle of the unconsciousness when humans try to solve hard mathematical problems. He explains this phenomenon by its most important feature, namely that it can make (and indeed makes) use of concurrent search (whereas conscious thought cannot be concurrent), see p. 22 Hadamard (1944): *"Therefore, we see that the unconscious has the important property of being manifold; several and probably many things can and do occur in it simultaneously. This contrasts with the conscious ego which is unique. We also see that this multiplicity of the unconscious enables it to carry out a work of synthesis."* That is, in Hadamard's view, it is important to follow different lines of reasoning simultaneously in order to come to a successful synthesis.

Human reasoning has been described in traditional AI (e.g., expert systems) as a process of applying rules to a working memory of facts in a recognise-act cycle. In each cycle one applicable rule is selected and applied. While this is a successful and appropriate approximation for many tasks (in particular for well understood domains), it seems to have some limitations, which can be better captured by an approach that is not only cooperative but also concurrent. Minsky (1985) gives convincing arguments that the mind of a single person can and should be considered as a society of agents. Put in the context of mathematical reasoning this indicates that it is necessary to go beyond the traditional picture of a single reasoner acting on a working memory – even for adequately describing the reasoning process of a single human mathematician.

There are two major approaches to automated theorem proving, machine-oriented methods like the resolution method (with all its ramifications) and human-oriented methods. Most prominent amongst the human-oriented methods is the proof planning approach first introduced by Bundy (1988). In our poster we argue that an integration of the two approaches and the simultaneous pursuit of different lines in a proof can be very beneficial. One way of integrating the approaches is to consider a reasoner as a collection of specialised problem solvers, in which machine-oriented methods and planning play different rôles.

2 System Architecture

The architecture (for further details see Benzmüller et al. (1999)) that we describe here allows a number of proof search attempts to be executed in parallel. Each specialised subsystem may try a different proof strategy to find the proof of a conjecture. Hence, a number of different proof strategies are used at the same time in the proof search. However, following all the available strategies simultaneously would quickly consume the available system resources consisting of computation time and memory space. In order to prevent this, and furthermore, to guide the proof search we developed and employ a resource management concept in proof search. Resource management is a technique which distributes the available resources amongst the available subsystems (cf. Zilberstein (1995)). Periodically, it assesses the state of the proof search process, evaluates the progress, chooses a promising direction for further search and redistributes the available resources accordingly. If the current search direction becomes increasingly less promising then backtracking to the previous points in the search space is possible. Hence, only successful or promising proof attempts are allowed to continue searching for a proof. This process is repeated until a proof is found, or some other terminating condition is reached. An important aspect of our architecture is that in each evaluation phase the global proof state is updated, that is, promising partial proofs and especially solved subproblems are reported to a special plan server that maintains the progress of the overall proof search attempt. Furthermore, interesting results may be communicated between the subsystems (for instance, an open subproblem may be passed to a theorem prover that seems to be more appropriate). This communication is supported by the shells implemented around the specialised problem solvers. The resource management mechanism analyses the theorem and decides which subsystems, i.e., which provers, should be launched and what proportion of the resources needs to be assigned to a particular prover. The mechanism is also responsible for restricting the amount of information exchange between subsystems, so that not all of the resources are allocated to the communication. The Figure to the right demonstrates this concurrent resource management based proof planning architecture. The involved planning agents are represented by PA$_n$ and the ovals indicate the amount of resources assigned to them in each reasoning phase.


Theorem
Assessment / Evaluation (Communication/ Exchange of Results)
Partial Proof
PA1
PA2
PA3
PA4
PA5
PAk
PAm
PAn
Completed Proof

We argue that the effect of resource management leads to a less brittle search technique which we call focused search.

Breadth-first search is robust in the sense that it is impossible to miss a solution. However, it is normally prohibitively expensive. Heuristic search may be considered as the other extreme case, it is possible to go with modest resources very deep in a search tree. However, the search is brittle in that a single wrong decision may make it go astray and miss a solution, independently of how big the allocated resources are. Focused search can be considered as a compromise — it requires more resources than heuristic search, but not as much as breadth-first search. As a result, a solution can still be found even if the focus of the search is misplaced. Clearly, more resources are necessary in the case of a bad than of a good focus.

We currently realise the so-called focused proof search as an adaptation of the multi-agent planning architecture, MPA Wilkins and Myers (1998), in the proof planning domain. Important infrastructure for this enterprise is provided by the ΩMEGA (http://www.ags.uni-sb.de/~omega/) proof development environment. The main component of MPA is a multi-agent proof planning cell, which consists of 1) several planning agents, 2) a plan server, 3) a domain server, and finally 4) a planning cell manager.

1. The quite heterogeneous reasoning systems (first-order reasoners, higher-order reasoners, computer algebra systems, etc.) already integrated to ΩMEGA are available as planning agents. An interactive user may become a concurrent planning agent as well.

2. The plan server stores promising partial proof plans returned by the planning agents in their previous runs within a unified data format. This enables backtracking on two distinct levels: we can backtrack within the actual proof plan by taking back single proof steps or subproofs contributed by some of the planning agents, and we can completely shift to some alternative proof attempt that has been abandoned previously.

3. A domain server provides the necessary knowledge for the planning cell manager as well as for the single planning agents. In our context it consists of a structured database of mathematical theories. Moreover, it should contain domain specific knowledge relevant to certain planning agents.

4. The planning cell manager re-organises and controls the reasoning process in each iteration phase based on its (and/or the user's) crucial evaluation and assessment considerations. Its prototype is based on the agent-architecture described in Benzmüller and Sorge (1999) allowing for a close and flexible integration of an interactive user into automated reasoning processes.

3 Conclusion

Our work does not directly follow the long-term goal of building a 'complete mind'. However we think that we will encounter many of the problems in our limited domain which will have to be solved in building a complete mind. In particular a distinction between different levels, reactive and deliberative modes, meta-level reasoning and so on, seems to be very important in the wider context of mathematical reasoning.

References

C. Benzmüller and V. Sorge. Critical Agents Supporting Interactive Theorem Proving. *Proceedings of EPIA-99*, Volume 1695 of *LNAI*, 1999. Springer.

C. Benzmüller, M. Jamnik, and M. Kerber a nd V. Sorge. Towards concurrent resource managed deduction. Tech-Report CSRP-99-17, The University of Birmingham, School of Computer Science, 1999.

A. Bundy. The Use of Explicit Plans to Guide Inductive Proofs. *Proceedings of the CADE-9*, volume 310 of *LNCS*, 1988. Springer Verlag, Berlin, Germany.

J. Hadamard. *The Psychology of Invention in the Mathematical Field.* Dover Publications, New York, USA; edition 1949, 1944.

M. Minsky. *The Society of Mind.* Simon & Schuster, New York, USA, 1985.

D. E. Wilkins and K. L. Myers. A Multiagent Planning Architecture. *Proceedingsof AIPS'98*, 1998. AAAI Press, Menlo Park, CA, USA.

S. Zilberstein. Models of Bounded Rationality. In *AAAI Fall Symposium on Rational Agency*, Cambridge, Massachusetts, November 1995.

Architecture of Mind Considering Integration of Genetic, Neural, and Hormonal System

Stevo Bozinovski[1], Liljana Bozinovska[2]
[1]Electrical Engineering Department, University of Skopje, Skopje, Macedonia
[2]Physiology Institute, Medical Department, University of Skopje, Skopje, Macedonia
bozinovs@rea.etf.ukim.edu.mk

Abstract

The problem of building an architecture of a mind which appreciates three crucial systems from biology: genetical, neural, and hormonal systems, is considered. It is presented a generic architecture and a derivate, an emotion learning architecture. A learning rule which explicitly implements the influence of the mentioned three systems is proposed.

1. Introduction: Problem Statement

Although in earlier stage most AI approaches insisted on symbolic reasoning with no particular reference to biology, most contemporary AI approaches show interest to concepts as artificial neural networks and genetic algorithms, evidently motivated by biology. However, biological agents exhibit behavior that is also influenced by the hormonal system. That motivates statement of the following problem:

Find architecture of mind that will implement neural, genetic, and hormonal control.

Symbolically, we need agent architecture with the following control function:

Behavior = Control(Neural, Genetic, Hormonal)

That problem we call the problem of integrated biology-inspired (IBI) control. In particular we are interested in *learning architectures* with a property of IBI control.

2. A Conceptual Architecture

Our approach toward IBI architecture for an agent is shown in Figure 1. As Figure 1 shows, the agent, from the *genetic environment*, inherits initial states of its memory and also other initial parameters for behaving in the *behavioral environment.* The main control system, the neural system, controls the motor response of the agent, including the secretory response of the glands. It controls the hormonal system, which in turn can produce emotions, moods, and other states of the consciousness that affect the nervous system. Through the behavioral environment interface the agent interacts with the behavioral environment. The operating system supplies features such as priorities, preferences, goals, needs, queues, activation strategies, thresholds, among other parameters and functions required for cooperation between the mentioned systems within an agent.

The agent can learn and adapt to a changing environment. The agent is able to import and export genomes, data structures reflecting the adaptation of the agent in the considered environment.

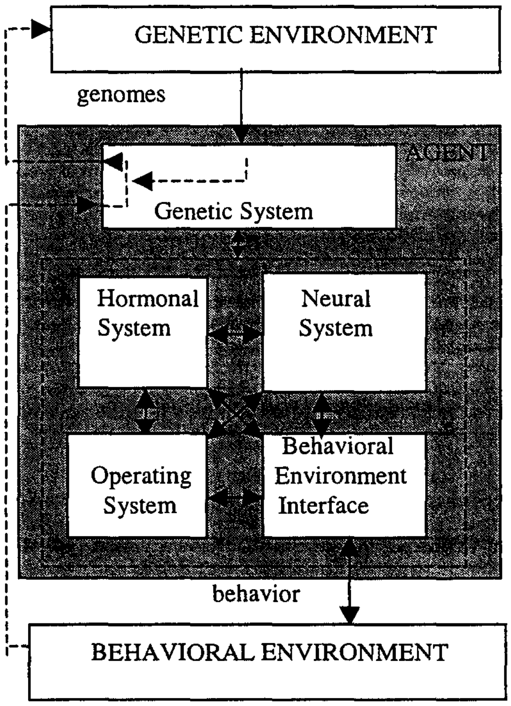


Figure 1. A generic IBI control architecture

In the genetic environment, genomes are transferred to other agents, to speed up the adaptation of the new generation of agents in the behavioral environment.

3. A Working Architecture

Figure 2 shows an architecture, which can be viewed as derived from the architecture shown in Figure 1. In this IBI architecture instance, the neural system is represented by a crossbar connected neural weights matrix, the hormonal system influences emotions, the behavior environment interface receives situations and computes actions, while the operating system supplies only some personality parameters, such as curiosity to action selection and sensitivity threshold to emotion computation.

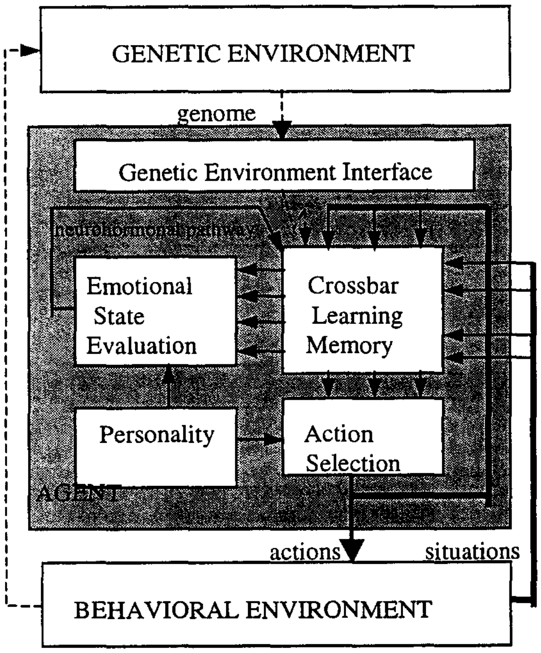


Figure 2. The CAA agent architecture

A crossbar computing procedure over the weights matrix is used for computing emotions (column-wise) and actions (row-wise). This architecture we call Crossbar Adaptive Array (CAA) architecture.

4. Emotion Learning

It is assumed that each crossbar element, w_{aj}, represents an emotion, *Emotion(a,j),* of performing action *a* in situation *j*. Having that, CAA performs its *crossbar emotion learning procedure,* which has four steps:

1) state *j*: choose an action in situation: (let it be action *a*; let the environment returns situation *k*)
2) state *k*: *feel the emotion* for state *k*: *emotion(k)*
3) state *j*: *learn the emotion* for *a* in *j*: *Emotion(a,j)*
4) change state: *j=k*; goto 1

This learning procedure is an emotion backpropagation procedure (secondary reinforcement learning procedure). The learning rule used in CAA in step 3), is

$$Emotion^{0}(a,j) = genome^{0}(envir) \qquad (1a)$$
$$Emotion'(a,j) = Emotion(a,j) + emotion(k) \qquad (1b)$$

It is a simple learning rule, which just adds the *emotion of being* in the consequence situation, *k*, to the *emotion toward* performing action *a* in situation *j* on which *k* is the consequence.

5. Related work

In this short paper we presented some conceptual issues related to the CAA architecture. Implementations are described in Bozinovski (1999), and Bozinovski et. al. (1999). The work presented here is related to contemporary reinforcement learning research (see Barto, 1997) and contemporary emotion research (see Castelfranchi, 2000).

6. Conclusion

The learning rule (1) includes influence from the genetic environment, which is assumed to reflect the behavioral environment, and performs emotion learning, where *emotions are signaled by the hormonal system and are stored in the neural system.*
Symbolically, we can rewrite (1) as

$$neural^{0}(a,j) = genetic^{0}(envir) \qquad (2a)$$
$$neural'(a,j) = neural(a,j) + neurohormonal(k) \qquad (2b)$$

It is a learning rule of an IBI architecture we were searched for.

References

A. Barto. Reinforcement learning. In O. Omidvar and D. Elliot (Eds.) Neural Systems for Control, pp. 7-29, Academic Press 1997

S. Bozinovski. Crossbar Adaptive Array: The first connectionist network that solved the delayed reinforcement learning problem. In A. Dobnikar, N. Steele, D. Pearson, R. Albrecht (Eds.) Artificial Neural Nets and Genetic Algorithms pp. 320-325, Springer, 1999

S. Bozinovski, H. Jaeger, P. Schoel. Engineering goalkeeper behavior using an emotion learning method. In S. Sablatnoeg, S. Enderle (Eds.) Proc RoboCup Workshop, KI99: Deutsche Jahrestagung fuer Kuenstliche Intelligenz, pp. 48-56, Bonn, 1999

C. Castelfranchi. Affective appraisal vs cognitive evaluation in social emotions and interactions. In A. Paiva, C. Martinho (Eds.) Affect in Interactions, Springer, 2000

Can we model the mind with its own products?

Marcin Chady
The University of Birmingham, UK
M.Chady@cs.bham.ac.uk

Abstract

This poster is an invitation to a free discussion on the limitations of the mind in its capability to understand itself and other minds. It poses some questions, presents examples and makes suggestions, in the hope to attract some comments and feedback from the readers.

The human brain has evolved to model the world, so that we can avoid dangers and achieve our goals. It is not possible to simulate the universe exactly, so some (arguably significant) degree of approximation is required. In order to draw conclusions and predict future events, one must generalise, so any brain-endowed organism is a natural classifier, and we humans are a prime example of this: we categorise, generalise and organise things into hierarchies, so that we can make better sense of the world, and turn its rules into our advantage. We have developed languages, algebras and calculi. Our sophisticated system of symbols can describe abstract as well as physical things. However, as we face the enigma of our own minds, we must ask ourselves the question: Can we model the mind with its own products?

When we talk about thinking, we tend to use inherently ambiguous terms: emotions, intentions, exploration, control, or even memory. It isn't until we actually try to realise them in a physical system, that it transpires how vague they are. Many of them have been discredited from scientific discourse, and confined to the realms of folk psychology, but much of the confusion remains (Smolensky (1988)). People like to view the mind as a number of parallel processes: transforming information, exchanging signals, making decisions, each specialised in a different kind of operations. It is easy to conceptualise such systems using the "divide and conquer" strategy, and, perhaps most importantly, it is easy to *depict* them using boxes and arrows.

Although symbolic systems allow a certain amount of flexibility and/or indeterminism in their architecture (cf. the ACT-R system or Sloman's models of emotional agents, see e.g. Sloman (1999)), the component-based perspective still remains. It is easy to forget that the functionality and the mechanism are not necessarily in a one to one relationship. Consequently, we tend to make statements like "if there are two motive generators competing, then there must be a motive comparator". Such reasoning assumes that each function identified in the designer's mind has an architectural embodiment. Yet, it is easy to imagine competition without an arbiter, in which case the "comparator" has no counterpart. And what about all those possibilities which are not so easy, or even impossible to imagine?

Much of the non-symbolic AI, such as neural networks, inadvertently follows the same path by giving parts of their systems explicitly defined functionality, and combining them into interacting ensembles of modules. A prime example of this are implementationalist systems, where the logic is hand-wired into the connections between units (e.g. Shastri and Ajjanagadde (1993), Barnden (1991)). Not only do they risk missing out a significant part of the solution domain, but parts of the problem domain too: they usually end up with a limited capability for learning.

It is hard not to structuralise, given that this is what our brains have been designed to do through millennia of evolution. However, one must not forget that everything our cognitive processes come up with is an artefact. It's a product of a process whose purpose is by no means general. On the contrary, its purpose is to filter out details which are irrelevant to survival, and produce a *convenient* model of the environment. There is no reason why our cognitive apparatus should be able to cope with its own workings, just like no one is predisposed to imagine 7-dimensional objects, which some physicists suggest our universe really consists of.

Can we escape the kaleidoscope of our cognition? Most probably not. Even the machines we create will be biased. However, we can reduce this bias by limiting our involvement in their design, putting emphasis on self-organisation rather than designing complete architectural solutions. Evolutionary strategies spring to mind. But we can also study complex unconventional systems and look for new insights there. Of course, every insight will be subjective too, but there is nothing we can do about it. At least we may discover a new way of thinking which will take us a further step back from the narrow vision of a "survival machine".

An example of such approach can be found in the work of Hanson and Crutchfield (1992). There, a complex behaviour of CA is analysed using classic paradigms from computational theory, such as Finite State Automata. The work demonstrates how complicated operation can result from simple interactions between simple units. Conversely, the use of classical computational models to describe the global behaviour shows that no explicit FSA machinery is necessary to produce FSA behaviour.

Perhaps then, there are no parts responsible for motivation, emotions, control, etc. They could be just facets of the same process, presenting themselves differently from different angles. Much like the facets of a hypercube intersected with a 3-dimensional hyperplane, or like ephemeral vortices of turbulent flow. And, last but not least, let us not forget that the mind itself is an artefact: an attractive and convenient concept for picturing our thoughts perhaps, but not necessarily the most useful one in tackling the "hard problem".

References

John A. Barnden. Encoding complex symbolic data structures with some unusual connectionist techniques. In John A. Barnden and J. B. Pollack, editors, *High Level Connectionist Models*, volume 1 of *Advances in Connectionist and Neural Computation Theory*. Ablex, Norwood, N.J., 1991.

Jim Hanson and Jim Crutchfield. The attractor-basin portrait of a cellular automaton. *Journal of Statistical Physics*, 66:1415–1462, 1992.

Lokendra Shastri and Venkat Ajjanagadde. From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, 16:417–494, 1993.

Aaron Sloman. What sort of architecture is required for a human-like agent? In Michael Wooldridge and Anand Rao, editors, *Foundations of Rational Agency*. Kluver Academic, Dordrecht, 1999.

Paul Smolensky. On the proper treatment of connectionism. *The Behavioral and Brain Sciences*, 11, 1988.

Towards an Axiomatic Theory of Consciousness

Jim Cunningham, Imperial College, London SW7 2BZ

rjc@doc.ic.ac.uk, March 2000

Pragmatics of Consciousness

Although part of the ancient mind-body problem of philosophy, the concept of consciousness itself is well enough recognised for it to be an ordinary word of our language. A conscious individual is aware, and knowing; the unconscious condition is normally recognisable. Yet numerous popular and contemporary books by Searle, Dennett, and others, show its explication to be contentious and a challenge to our suppositions on reality; a hazardous topic indeed for a would-be engineer of artificial intelligence.

Our justification for addressing the subject is that artificial agents which display elements of intelligent behaviour already exist, in the popular sense of these words, but that we would doubt the intelligence of an agent which seemed to us to have no sense of "self", or awareness of its capabilities and its senses and their current state. So an approximation to human consciousness could enable us to converse more naturally with an individual agent. We do need an account for the first person perspective as well as the second and third, and we cannot rule out the possibility that consciousness has utilitarian function, evolved to ensure survival.

Contention arises over whether consciousness can be considered a mental state of the human mind, for this brings presuppositions of the intentional stance and issues of faithfulness to the human model. But lack of faithfulness to a biological model is not a barrier to engineering, as the wheel, the fixed wing, and the computer itself demonstrate. Software agents are already designed with mental states and practical reasoning methods which have emerged as abstractions from rational enquiry rather than any physical brain model. While agent designers may also eschew such models, and instead rely on a variety of physical and computational devices, in well known cases the management of complexity leads to design architectures with layers of abstraction, some of which are comparable with intentional models of the mind.

To bypass the metaphysics of consciousness in favour of pragmatic considerations, there is evidence to consider, the view of peers in rational enquiry, and the need for guidance in an artificial construction. Clinical reports, psychological experiment, and philosophical enquiry, lead to a variety of theories which partially explain the phenomena and suggest layers of consciousness. Problematic issues range from neurological phenomen such as phantom limbs and the relation to wakefulness and unconscious mental processing, through issues of identity and the effect of emotion and the habitual, to an explication of context and presence in perception and of the links with language and intentionality.

Our formalism arises from attempts to bridge the gap between agents designed with mental states, and credible multi-processing implementations. It may be compatible with the implementations of a psychologically motivated theory like that of Baars, which can be realised as a computational agent with a myriad of heterogenious processes. But to explicate conscious behaviour we may still require layers of conception which we hardly discuss here.

A Refined Intentional Stance

Mental models of the intentional stance encroach on two areas of agent design. One is as an abstract basis for incorporating plans and the selection of actions through means-end reasoning in software agents, notably in variants of Rao and Georgeff's Belief, Desire, Intention (BDI) paradigm. The other area is the related basis for giving definition to standard acts of communication as realisations of speech act theory, so that there are ingredients of a coherent basis for dialogue between agents in terms of what we can loosely call knowledge interchange.

Our proposal for steps towards an axiomatisation of consciousness depends critically on a refinement of traditional ideas of intentionality. From the perspective of an agent designer, extant intentional theories of rational agents focus on *stative* concepts like the BDI concepts themselves, and *knowledge* and *commitment*, each of which can be regarded as expressing computational *data* states. Activity, or process states, which are equally important in a computational model, have been ignored, or rather buried in naive computational models. But activity states like *planning*, *learning* and *sleeping*, and the *sensing* and *perceiving* of external conditions are equally important for a computational model of a rationality. This is a serious deficiency in the usual perception of

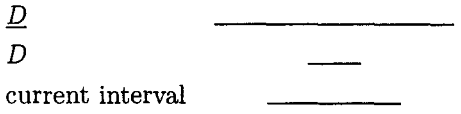


Figure 1: interval relations D and $\underline{D}$

mental state, and we suggest a remedy below. However, there is another defect. The usual axiomisation for belief, and of knowledge, presumes introspection; e.g. for knowledge, that which is known is known, that which is not known is known to be not known. These are strong conditions which make such states already too "conscious" for some forms of memory recall and learnt behaviour.

The limitations of stative mental states can be overcome by simply allowing activity states as well. Each class can be considered durative on a temporal frame, but we gloss over the homogeneous / compositional distinction by using the progressive form of the activity. We capture this by a modal operator *prog* to modify the singular verb predicate, so that, for example, a rendering of *j is sensing c* becomes $prog\ senses_j\ c$. To define the *prog* operator we use the interval temporal logic of Halpern and Shoham, which uses modal operators to incorporate the interval relations identified by Allen. The during relation $\mathbf{D}$, and its complement $\underline{\mathbf{D}}$ with respect to the current interval are illustrated in figure 1. We may define *prog p* to be the coercing form $\langle\underline{D}\rangle[D]p$, read (right to left) as *p holds on all sub-intervals during some interval which contains the current interval.* The usefulness of the Halpern and Shoham logic for representing and modelling both aspect and tense is demonstrated in a thesis by Leith and is being explored in joint work with Lloyd Kamara.

Once we have the ability to express temporal relations between interval based activities as logical properties, the interactions between activities and other mental states can be expressed by axioms. We may for instance consider that an axiom like:

$$\begin{aligned} &prog\ perceives_j\ p \leftrightarrow \\ &\quad prog\ senses_j\ c\ \wedge\ prog\ remembers_j\ (c \rightarrow p) \end{aligned}$$

expresses the idea that sensory perception amounts to an ongoing inferential interaction between autonomous sense and memory recall processes. Although we could be more precise by constructing a compound sense process from realisations of particular sense mechanisms, the more problematic element of this definition is memory recall itself. Here we postulate a subconscious process rather than a stative belief state, because as mentioned above this would already be too strong a condition. It seems too that we have capability for richer expressions of tense and aspect which could encapsulate aspects of learning.

Introspective Awareness

The perception processes posited above could be those of a sophisticated but unconscious automaton. The awareness needed for consciousness can also be considered a mental process rather than a data state, one which includes sensory perception, but also meta perception of sensory perception. Awareness can also be switched on and off by paying attention, either in response to change in sensory perception, or through volition; primitive processes whereby mental activity and ultimately action are controlled. We follow Carl Ginet by arguing not only for such philosophical abstractions but because mental control processes relate to notions of will and causality through neurological elements such as the motor cortex.

Thus to introduce consciousness, and ultimately a consciousness of responsibility, we need the activity of being aware to be positively introspective and controllable to some degree. When an agent is aware, it not only perceives, but at least for some conditions, perceives that it perceives. Thus assuming positive meta perception only, we might provide an axiom for a progressive form of introspective awareness as:

$$\begin{aligned} &prog\ aware_j\ p \leftrightarrow prog\ perceives_j\ p \\ &\qquad \wedge\ prog\ perceives_j\ prog\ perceives_j\ p \end{aligned}$$

This may be unnecessarily clumsy, not strong enough, and obscuring distinct perception processes, but because the scope and degree of introspection can be graded there seems to be no evolutionary argument against the acquisition of such higher levels of perception, indeed it seems necessary for a sense of social responsibility. A socially conscious agent which perceives a causal relationship will also perceive consequences of its perception of this relationship. We claim that once an agent has mental activities of sufficiently introspective awareness it also has a form of consciousness, that in its weakest form consciousness is just a progressive activity of being introspectively aware of something: $\exists p.prog\ aware_j\ p$. Graded and focused consciousness can follow.

Bibliography

B.J.Baars,*In the theater of consciousness*, OUP1997
M.Bratman,*Intention, Claims and Practical Reason*, Harvard University Press 1997
D.Dennett, *Consciousness Explained*, Penguin 1991
C.Ginet, On Action, CUPress 1990
M.F.Leith, "Modelling Linguistic Events", PhD Thesis, Imperial College, University of London 1997
A.S.Rao and M.P.Georgeff,"BDI agents:from theory to practice", ICMAS-95, San Francisco, CA, 1995
J.Searle, *Mind, Language and Society*, 1999 (Orion)

Design Issues of Biological and Robotic 'Minds'

Kerstin Dautenhahn
Department of Cybernetics, University of Reading, United Kingdom
Department of Computer Science, University of Hertfordshire, United Kingdom (from 1/4/2000)
K.Dautenhahn@cyber.rdg.ac.uk, K.Dautenhahn@herts.ac.uk (from 10/4/2000)

Abstract

This abstract discusses the social dimension of biological and robotic 'minds'. Firstly, the evolutionary perspective is addressed. The *Narrative Intelligence Hypothesis* suggests that the evolutionary origin of communicating in stories was correlated with increasing social dynamics among our human ancestors. Secondly, it is suggested that a variety of different 'minds' exist and have evolved, e.g. the autistic mind. These issues are related to current projects on social robots which the author is involved in. Implications for designing a functioning mind are discussed.

1 What Are Minds For? The Evolutionary Perspective

Designing a functioning mind can benefit from analysing the conditions and constraints which have shaped the evolution of animal minds. Minds are certainly attributed to members of *Homo sapiens* (and as some evidence suggests several other hominid species might have existed with 'minds'), but other candidates exist among mammals (e.g. non-human apes, dolphins, elephants) and birds (e.g. parrots and members of the crow family). Interestingly, species which we describe as possessing a 'mind' are usually highly social. Even the 'solitary' life style of *Pongo pygmaeus* (orangutan) (who nevertheless seem to be highly social in their ability to recognise and interact with each other) is rather a secondary adaptation to a particular environment which demands a spatially distributed social organisation. The *Social Intelligence Hypothesis* suggests that primate intelligence primarily evolved in adaptation to social complexity, i.e. in order to interpret, predict and manipulate conspecifics (Byrne and Whiten, 1988). However, as Richard Byrne recently pointed out (Byrne, 1997), this hypothesis might account for the evolution of *primate* intelligence, but *not* for the specific *human* kind of intelligence. This suggests that other factors (e.g. language as suggested by Byrne (1997) and others) played a significant role in the evolution of human intelligence. If language was a major milestone in human evolution, what attributes of language made it superior to other forms of communication?

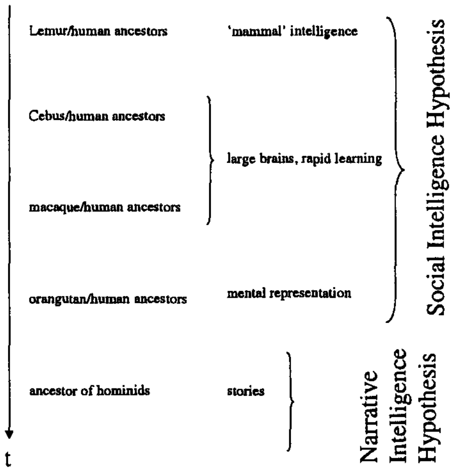


Figure 1: Evolution of the Social Mind, modified from Byrne (1997).

In (Dautenhahn, 1999) I discuss how narrative psychology and studies on the development of autobiographic memory and a 'self' give evidence which suggests that 'stories' are the most efficient and natural human way to communicate, in particular to communicate about others (Bruner, 1991). The *Narrative Intelligence Hypothesis* (Dautenhahn, 1999) proposes that the evolutionary origin of communicating in stories was correlated with increasing social dynamics among our human ancestors, in particular the necessity to communicate about third-party relationships (which in humans reaches the highest degree of sophistication among all apes, e.g. gossip), see Fig. 1. The lessons for designing a functioning mind are: a) minds need to be designed as social minds, b) a human-style social mind need to be able to communicate in 'stories'.

2 The Case of Autism: Diversity and Adaptive Radiation of 'Minds'

Autism is a specific disorder which results in significant deficits in the social domain: people with autism generally have great difficulty relating to other people, interacting socially in an appropriate way (Baron-Cohen, 1995). Rather than considering people with autism as having 'defective' minds, they can be viewed as possessing minds which are functioning but *different* from other people. Similarly, other non-human animals might possess minds equally 'powerful' as ours, (see Herman (2000) for a comprehensive assessment of dolphin intelligence), but different, and often difficult to study due to our limited understanding of the species and their environments. Natural evolution supported diversity and adaptive radiation so that different minds might have evolved in adaptation to particular environmental constraints (biotic and abiotic) and thus creating a particular niche. A 'general purpose animal mind' does not exist. Lessons and speculations for designing functioning minds: 1) A single architecture as a 'solution' to designing a functioning mind is unlikely. The design spaces of natural and artificial minds are still to be discovered, but we can expect a high degree of diversification. 2) For artificial minds, a number of constraints (e.g. body shape of robots) are under our (the designer's) control and this synthetic approach could further our understanding of minds complementary to investigating animal minds. 3) In the same way as the notion of 'fitness landscape' helps biologists in understanding evolution and speciation, a similar concept might be developed in order to describe and evaluate the 'fitness' of artificial systems.

3 Robotic Friends

The project AURORA investigates how an autonomous robotic platform can be developed as a remedial tool for children with autism (Dautenhahn 1999, Dautenhahn & Werry 1999). A deliberately non-humanoid robot with a simple behaviour repertoire is used as a toy, providing an 'entertaining' and playful context for the autistic child in which it can practise interactions (specific issues addressed are e.g. attention span and eye contact). The project poses several challenges for developing an artificial mind: 1) how can the robot's mind 'grow'?, 2) how can narrativity develop, what are its precursors?, 3) the role of physical interaction and embodiment: what can the physical robot provide that cannot be studied in using software systems?, 4) can the robot ultimately serve as a 'mediator' between the child and the teacher and/or other children? 5) what is the relationship between the internal control architecture (the 'artificial mind') and the way the robot behaves believably and 'mindfully'?

Acknowledgements

The AURORA project is supported by EPSRC (GR/M62648), *Applied AI Systems Inc.*, and the National Autistic Society (NAS).

References

AURORA:http://www.cyber.rdg.ac.uk/people/kd/WWW/aurora.html

S. Baron-Cohen. *Mindblindness*. MIT Press, 1995.

J. Bruner. The Narrative Construction of Reality, *Critical Inquiry,* 18(1):1-21, 1991.

R.W. Byrne and A. Whiten. *Machiavellian Intelligence.* Oxford: Clarendon Press. 1988.

R.W. Byrne. Machiavellian Intelligence. *Evolutionary Anthropology*, 5:172-180, 1997.

K. Dautenhahn. Robots as Social Actors: AURORA and The Case of Autism, *Proc. CT99* pp. 359-374, 1999.

K. Dautenhahn. The Lemur's Tale - Story-Telling in Primates and Other Socially Intelligent Agents. Proc. *Narrative Intelligence*, AAAI FS 1999, Technical Report FS-99-01, pp. 59-66, 1999.

L. M. Herman. Vocal, Social, and Self Imitation by Bottlenosed Dolphins. *Imitation in Animals and Artifacts,* K. Dautenhahn and C. L. Nehaniv (Eds.), in press.

I. Werry and K. Dautenhahn. Applying robot technology to the rehabilitation of autistic children. *Proc. SIRS99*, pp. 265-272, 1999.

The Baby Webmind Project

Ben Goertzel, Ken Silverman, Cate Hartley, Stephan Bugaj, Mike Ross
Intelligenesis Corp.; 50 Broadway, NY NY 10004
ben@intelligenesis.net, ken@intelligenesis.net, cate@intelligenesis.net,
stephan@intelligenesis.net, miro@intelligenesis.net

Abstract

The Baby Webmind project involves teaching an instance of the Webmind AI system how to perceive, act and cognize through interaction in a shared perceptual environment. This talk describes the goals and methodology of the project at a high level, and briefly reviews some of the AI technologies underlying it.

1 Introduction

Webmind is an original AI architecture based largely on the interdisciplinary model of the mind developed in Goertzel (1994, 1997). It is based on a vision of the mind as an evolutionary, self-organizing, self-producing system. It is integrative in nature, involving reasoning, perception, cognition, evolution, long and short term memory and other modules embedded in a single dynamic data structure which allows emergent learning amongst the different components.

The goal of the project is not to simulate a human mind or to create a program that can pass the Turing test, but rather to create a software system with a mind. Our working definition of a mind is: a collection of patterns that forms and perceives patterns in itself and the world, in order to achieve complex goals in a complex environment. In Webmind's case the complex environment is the Internet.

The process of engineering Webmind has been ongoing since late 1997, and has now, thanks to the efforts of more than 40 AI engineers, reached a point where the major components of the system are complete and the focus can be placed on the integration of components and the induction of emergent intelligence in the whole system. This involves, among other things, a process of educating Webmind by interactive learning in a shared environment what we call "bringing up Baby Webmind."

2 Webmind

Webmind is a massively parallel network of static and dynamic information agents that continually recompute their relationships to other agents, implemented as a distributed Java software system.

The information agents populating the massively parallel self-organizing network that is Webmind come in many different species, each one specialized for different purposes. And all the different kinds of agents can learn from each other – the real intelligence of Webmind lies in the dynamic knowledge that emerges from the interactions of different species of agents.

There are numerical data processing agents, that recognize patterns in tables of numbers, using a variety of standard and innovative algorithms.

There are text processing agents, that recognize key features and concepts in text, drawing relationships between texts and other texts, between texts and people, between texts and numerical data sets.

There are reading agents, which study important texts in detail, proceeding through each text slowly, building a mental model of the relationships in the text just like a human reader does.

There are textual-numerical correlation agents, that recognize patterns joining texts and numerical data files together. These underly Webmind Market Predictor's unprecedented ability to find the concepts in news that drive the financial markets.

There are categorization agents of various kinds, that study the other agents in the mind, group them together according to measures of association, and form new agents representing these groupings.

There are learning agents, that recognize subtle patterns among other agents, and embody these as new agents. Among these are agents carrying out logical inference, according to a form of probabilistic logic based on Pei Wang's Non-Axiomatic Reasoning System (Wang, 19XX); and agents carrying out evolutionary learning, according to genetic programming (Koza, 1992), a simulation of the way species reproduce and evolve.

There are agents that model users' minds, observing what users do, and recording and learning from this information. There are agents that moderate specific interactions with users, such as conversations, or interactions on a graphical user interface. And there are self agents, that help Webmind study its own structure and dynamics, and set and pursue its own goals.

Each of these agents, in itself, has a small amount of intelligence, similar in some cases to that of competing AI products. The Webmind architecture provides a platform in which they can all work together, learning from each other and rebuilding each other, creating an intelligence in the whole that is vastly greater than the sum of the intelligences of the parts.

3 Teaching Baby Webmind

Even with all these diverse capabilities, the Webmind we have today is only a baby, exploiting 10architecture. To get all the mind modules to work together really intelligently, we need to lead the system step by step through goals, beginning with simple goals and gradually moving to more complex ones. We need to teach the system step by step almost like a human child.

Each of Webmind's modules is best of breed in some particular area, and the modules can be used independently or in various combinations to support various product functions. Putting a few modules together can give you functions that normal AI software can't do things like using text to predict the markets, or effectively filtering news messages for relevance. But putting all the modules together can get you actual intelligence, because the modules are chosen specifically so as to allow the system to understand itself, to recognize patterns in itself. To teach our baby Webmind, we plan to chat with it on a simple graphical/textual user interface. We won't chat with it about trees and flowers and teeth, because it doesn't have direct experience of these things. We'll chat with it about data files and shapes and MIDI music files, because these are the things that we can both experience. Intelligence has to be gained through interactive experience in a shared environment.

It's intriguing to see how the basic task of learning to interact in the world uses all Webmind's specialized modules. Reasoning and genetic programming evolution are used to find schema – sets of basic procedures for seeing and doing and thinking that are useful at achieving the system's goals and hence make the system happy. Categorization is needed to define contexts in the world a schema has to be judged by how it achieves important goals in relevant contexts. Language processing is obviously needed to chat with humans, and although in this context most of the specific nature of human language must be learned, nevertheless the basic structures needed for language understanding need to be provided from the start. Data processing is needed to turn raw numerical data files, sensed by the system, into comprehensible perceptual features. And so on. All the pattern finding and relationship building methods of Webmind's various modules are needed to provide the data that basic behavior schema need to act intelligently.

Acknowledgements

The ideas underlying this work owe a great deal to other members of the Intelligenesis AI staff, including especially Pei Wang, Jeff Pressing, Karin Verspoor, Anton Kolonin and Cassio Pennachin.

References

Ben Goertzel. *Chaotic Logic*. Plenum Press, New York, 1994.

Ben Goertzel. *From Complexity to Creativity*. Plenum Press, New York, 1997.

John Koza. *Genetic Programming*. MIT Press, Cambridge, MA, 1992.

Pei Wang. *Non-Axiomatic Reasoning System*. PhD thesis, Dept of CS & Cog. Sci. Indiana University, 1995.

Reflective Architectures for Survival: Bridging the Gap between Philosophy and Engineering

Catriona M. Kennedy
School of Computer Science
University of Birmingham
Edgbaston, Birmingham B15 2TT
Great Britain
C.M.Kennedy@cs.bham.ac.uk

Abstract

An essential feature of a mind is its ability to initiate original action in unforeseen circumstances. An artificial agent with this feature should recognise that its software is behaving differently in the new situation and take action if necessary to continue functioning as desired. This requires self-protection and creativity. We address the problem by exploring architectures for whole agents instead of looking for particular algorithms or specialist methods. This means that we are experimenting with patterns of interrelationships between specialist algorithms, without implementing the algorithms themselves in detail. To avoid excessive "shallowness" in the architecture, we use a bottom-up, low-level approach where an agent must "survive" in an environment in which its software can actually be damaged. Reflection enables the agent to monitor its internal execution patterns and detect any deviation from expectations. We are currently investigating distributed architectures where the functionality of an agent is produced collectively by two or more lower level agents which mutually observe and repair each other. First results indicate that the distributed architecture is more robust than a single-agent architecture.

1 Introduction

When we consider the mind's ability to cope with unusual events, there are many situations where it first recognises that there is something unusual about its own internal processing (e.g. a sudden apprehension or increased alertness) due to a still unidentified but novel feature of the environment. We find it is useful to model this as follows:

1. the unusual event is detected by the monitoring of *internal processes* (introspection), and not just by sensing the external environment, i.e. it is at least partly *reflective.*
2. this self-monitoring process is low-level and unconscious, but may be part of the micro-structure of the higher-level process by which we deliberately and consciously observe ourselves, which has been called meta-management (Sloman, 1997).

We are not claiming that this is exactly how a human mind works; instead we use it as a basis for the design of autonomous agents with self-protective capabilities in unpredictable and hostile environments. We believe that this problem must be solved if we are to understand the evolution of higher-level mental processes, which cannot exist as abstract entities, but must have a biological foundation. See e.g. Maturana and Varela(1980). Therefore, we first identify some essential features of a sufficiently challenging environment.

We define a *partially known environment* as one in which events can occur which are not taken into account by the agent's model of the world (based on current knowledge about it) and may include situations which the agent was not explicitly designed to handle. We call these events "anomalies".

We define a *hostile environment* as one in which the agent's executive and control systems do not cope well (e.g. they may be overloaded) or where they may be directly attacked; in other words, its software is subject to interference such as deception, modification or distraction (e.g. denial of service attacks). In these situations, the agent should recognise that its current software is not coping and either ask for help or find some innovative method of working around the problem.

2 Reflective Architectures

To investigate reflective architectures for autonomous agents, we focus on the comparison of whole architectures, instead of particular algorithms or techniques. We may characterise the architecture of an agent as the pattern of interrelationships (causal connections etc.) between entities representing specialist functions (e.g. they may be layers, individual components or sets of functionally similar components). Each entity may be regarded as a "slot" into which a specialist AI technique can be plugged in (e.g. a human interface layer or a planning specialist).

Since our question involves interrelationships, the "slots" of the architecture must be "shallow" (at least initially), otherwise the problem becomes unmanageable. In practise, this means that a slot contains only a minimal implementation of a technique.

However, an architecture that is too shallow may tell us nothing new. An extreme example is a synthetic "personality" that only displays a surprised facial expression in response to an external stimulus.

2.1 Situated reflection

We attempt to find a middle ground between shallow architectures and specialist techniques by ensuring that mentalistic concepts are given a concrete interpretation in an engineering sense. Therefore, we do not *simulate* a challenging environment, but ensure that it happens in reality, so that conditions are more similar to "survival" in the real world.

We implemented the requirements for a hostile, partially known environment as follows:

- *Hostile environment*: The agent is attacked at random intervals by a fault-insertion agent; faults may be produced in any part of the agent, including its self-monitoring and repair components.
- *Partially known environment*: The agent has no knowledge of the kinds of damage that can occur; instead it must detect anomalies in its software execution patterns as deviations from expected patterns. Then it must attempt to repair the damaged software or take evasive action as necessary.

To detect anomalies in its execution, it must have a model of its own normal operation, which is acquired gradually by self-observation during a protected "development phase". A simple type of model is that of a *signature* which is a collection of "normal" patterns of activity (Forrest et. al. 1994). We initially used a simple signature which required certain patterns to be present in execution traces.

2.2 Distributed reflection

The simplest form of reflection is a two-layered architecture containing a meta-level which monitors the agent's software execution patterns. The configuration where the meta-level is also applied to itself is shown schematically in figure 1(a). The whole agent is labelled A (the meta-level is not shown).

There are situations where the meta-level will not detect anomalies in itself (e.g. it cannot detect that it has just been deleted); in other situations it is unreliable (e.g. if the anomaly-detection process has been modified so that it gives false alarms and does not detect its own anomalous execution patterns). An alternative is a distributed architecture where the functionality of the first agent is produced collectively by two or more lower-level agents. Figure 1(b) shows this configuration for two agents (labelled A1 and A2). Figure 1(c) shows a possible three agent configuration. Each agent's meta-level not only monitors the agent's own software but also the meta-level of at least one other agent. In this way all meta-levels are protected. For related work, see Kornman(1996).

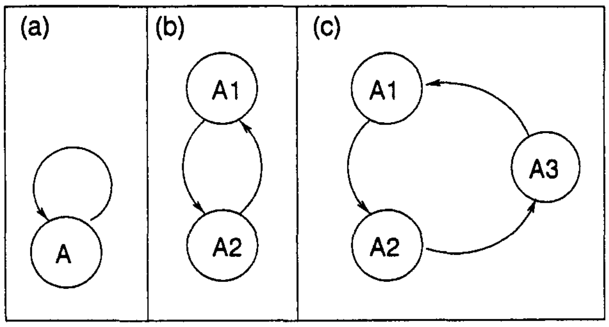


Figure 1: Reflective architectures

We are investigating the practical feasibility of a distributed architecture. Our first results indicate that it can provide survival advantages over a non-distributed version, but it requires a complex model acquisition process, involving higher-level category-formation. Details of the latest results can be found in Kennedy(2000).

References

S. Forrest et al. Self-Nonself Discrimination in a Computer. *Proceedings of the IEEE Symposium on Research in Security and Privacy*, 1994.

C. Kennedy. Reflective Architectures for Autonomous Agents. School of Computer Science Technical Report, April 2000.

S. Kornman. Infinite Regress with Self-Monitoring. Proceedings *Reflection'96*, San Fransico, April 21-23, 1996.

H. Maturana and F. J. Varela. *Autopoiesis and Cognition: The Realization of the Living*. D. Reidel Publishing Company, Dor.lrecht, 1980.

A. Sloman. What sort of architecture is required for a human-like agent? *Foundations of Rational Agency*, edited by Wooldridge, M. and Rao, A., Kluwer Academic Publishers, 1997.

LEARNING THE BASICS

Stefan Künzell
Justus-Liebig-Universität, Kugelberg 62, D-35394 Gießen
stefan.kuenzell@sport.uni-giessen.de

Abstract

The mind's basic task is to organize adaptive behaviour. It is argued that necessary conditions to achieve this are acquiring a 'body-self', a differentiated perception, motor intuition, and motor control. The latter three can be learned implicitly by crosswise comparing the perceived actual situation, the desired situation, the perceived result and the anticipated result.

1 Introduction

What is the functional role of a functioning mind? It is first and foremost designed to control behaviour in the most adequate way. This consideration implies that there cannot be a functioning mind without a body. So the starting point to design a functioning mind is to design a body with adequate action and perception. Speaking of "mind" instead of "brain" purports a certain potency of the behaviour control system. It should not be a hard-wired forward control system, like (more or less) an insect's brain, but an adaptable learning system. A functioning body-mind system needs to learn behaving flexibly in an ever-changing environment. Probability to "survive" increases if it predicts environmental changes correctly. This can only be done if it discriminates between what happens caused by the physics of the environment and what it causes to happen through its own action.

2 Learning tasks

Let us assume that designing a functioning mind depends only on adaptation starting at a *tabula rasa* state of mind. The only control mechanism available must be emotion, i.e. an evaluation system that provides the direction of learning. So the body-mind system's starting point is perceiving a stream of not interpretable noise and a feeling of discomfort.

2.1 Perception and the 'body-self'

One thing the body-mind system has to learn is to detect invariances in the stream of noise. The rating scale for the discrimination of invariances is the significance for its well-being. One significant invariance is for example the mother's face, her voice, the warmth of her skin, and the good feeling of being fed. One other significant invariance is that some entities in that noise persistently feed back a feeling when touched. They feed back pain when touched roughly, and warmth when touched tenderly.

Thus, perception (which is always directed) is being learned. And one of the first things being perceived is that some entities in the stream of noise belong physically to the body-mind system itself. It leads to a concept of a 'body-self'.

2.2 Motor intuition

The next thing the body-mind system has to learn is a mapping between the muscle commands, perceived environment and distal effects (e.g. Jordan and Rumelhart 1992), i.e. a forward model (for the engineer) or a motor intuition (for the psychologist). This is done by 'motor babbling'. Motor commands are produced in a random-like fashion. The invariant effects of the produced action (under environmental circumstances) are learned. This enables the body-mind system to anticipate its action's distal results, which enhances behavioural security (Hoffmann 1993) and provides a feeling of comfort or joy.

2.3 Motor control

Once it is able to anticipate the results, the body-mind system might "want" to produce them. I will not discuss the problem of the emergence of a "free will" here, that cause the desire. But admittedly it will be necessary to implement desires in some way for designing a functioning mind.

So the system has to learn the mapping between desired situation, perceived environment and motor behaviour, i.e. an inverse model (for the engineer) or motor control (for the psychologist). Jordan and Rumelhart (1992) developed a connectionist model for a small scale task in a static environment, where they integrated a forward and an inverse model for learning an controlling the movement of a two joint arm in a planar space.

3 Learning principles

In general, to enable learning, a body-mind system must have four concepts (implicitly) available in its mind: The perceived actual situation, the desired situation, the perceived true result available at the moment of the occurrence of the distal effect, and the anticipated result available at the moment of action. This implies the existence of an (implicit) memory, because the four concepts are not available in one time slot. For learning, the last three concepts are compared crosswise. We can distinguish four cases:

1. The true result equals the anticipated result, but both do not equal the desired situation. E.g. the system shoots a basketball to the basket, it fails, but in the moment of ball release it anticipates the failure. This is a usual case. Motor control, i.e. the inverse model has to be learned

2. The desired situation equals the anticipated result, but both do not equal the objective result. This is the case in novel situations. E.g. the system plays table tennis with always the same partner, which cannot play sliced balls. When a new partner now plays a slice, the system desires to return with a cross and in the moment of ball release it anticipates that the desired result will be achieved. But it does not; the perceived true result is that the ball leaves the bat in an unpredicted angle. In this case, perception must be differentiated. The environment's variance is mainly detected because the anticipated effect of a well-known action in an only seemingly well-known situation does not come true (see Hoffmann 1993 for further details).

3. If the desired situation equals the true result, but not the anticipated result, motor intuition must be learned. This is the case in trial and error learning, when suddenly, and not anticipated, action leads to the desired situation.

4. If all three concepts equal each other, everything is (presumably) fine and nothing must (can) be learned. This is the limit for implicit learning; improvement is only possible through presentation of explicit, consciously mediated knowledge of result.

4 Explicit vs. implicit learning

For implicit learning, the actual situation and the action's effect must be experienced. It is necessary to act. It is the privilege of self-conscious subjects to act cognitively instead of physically, to 'act as if you were acting'. A more or less correct motor intuition (or its conscious equivalent, motor imagery) and a concept of the 'body-self' presumed, distal results can be predicted mentally without acting. This protects consciously planning subjects from experiencing undesired or even lethal consequences, which enhances clearly the probability of survival of subjects and species.

To sum up, it is suggested here that for designing a functioning mind it is necessary to implement a functioning body-mind system, which is able to adapt to environmental changes without hardwired intelligence.

Acknowledgements

I thank the Cognition and Motor Behaviour Research Group (http://www.uni-giessen.de/~g51027/comoweb) for their helpful advices.

References

J. Hoffmann. *Vorhersage und Erkenntnis: Die Funktion von Antizipationen in der menschlichen Verhaltenssteuerung und Wahrnehmung.* [Prediction and Cognition: The function of anticipations in human behaviour and perception]. Hogrefe, Göttingen, 1993.

M. I. Jordan and D. Rumelhart. Forward Models: Supervised learning with a distal teacher. *Cognitive Science,* 16:307-354, 1992.

Of implementing neural epigenesis, reinforcement learning, and mental rehearsal in a mobile autonomous robot

Andrés Pérez-Uribe
Parallelism and Artificial Intelligence Group
Computer Science Institute, University of Fribourg
Chemin du Musee 3, CH-1700 Fribourg, Switzerland
Andres.PerezUribe@unifr.ch, http://www-iiuf.unifr.ch/pai/

Abstract

One of the key implications of functionalism is that minds can, in principle, be implemented with any physical substratum provided that the right functional relations are preserved. In this paper we present an architecture that implements neural epigenesis, reinforcement learning, and mental rehearsal, some of the functional building blocks that may enable us to build an artificial brain. However, we conclude that a new kind of machines, where the learning algorithms would emerge from the dynamics of the interconnection between the processing elements, are necessary for the implementation of cognitive abilities that are irreducible to a mechanistic computing algorithm.

1 Introduction

Based on the hypothesis that the physical matter underlying the mind is not at all special, and that what is special is how it is organized (Edelman, 1992), one come to the idea of building or simulating systems with functional capacities similar to those observed in nervous systems and brains to try to understand the mind.

From a biological point of view, it has been determined that the genome contains the formation rules that specify the outline of the nervous system. Nevertheless, there is growing evidence that nervous systems follow an environmentally-guided neural circuit building (neural epigenesis) (Sipper et al., 1997) that increases their learning flexibility and eliminates the heavy burden that nativism places on genetic mechanisms (Quartz and Sejnowski, 1997). The seminal work of the Nobel laureates D.H. Hubel and T.N. Wiesel on the brain's mechanism of vision (Hubel and Wiesel, 1979) describes a prime example of the role of experience in the formation of the neuro-ocular pathways.

The nervous system of living organisms thus represents a mixture of the innate and the acquired: "... the model of the world emerging during ontogeny is governed by innate predispositions of the brain to categorize and integrate the sensory world in certain ways. [However], the particular computational world model derived by a given individual is a function of the sensory exposure he is subjected to..." (LLinas and Pare, 1991).

Categorization, i.e., the process by which distinct entities are treated as equivalent, is considered one of the most fundamental cognitive activities because categorization allows us to understand and make predictions about objects and events in our world. This is essential in humans, for instance, to be able to handle the constantly changing activation of around 10^8 photo-receptors in each eye. Computational models of adaptive categorization have been developed and tested with success, and have been used to explain some sensory and cognitive processes in the brain such as perception, recognition, attention, and working memory (Grossberg, 1998). However, other types of learning, such as reinforcement learning, seem to govern spatial and motor skill acquisition (Sutton and Barto, 1998).

While in the former case only resonant states can drive new learning (i.e., when the current inputs sufficiently match the system's expectations) (Grossberg, 1998), in the latter "learning is driven by changes in the expectations about future salient events such as rewards and punishments" (Schultz et al., 1997).

2 Our neurocontroller architecture

We have developed a neurocontroller architecture (Fig. 1 based on the above premises (environmen-tally-guided neural circuit building for unsupervised adaptive clustering and trial-and-error learning of behaviors) and tested it using an autonomous mobile robot in a navigation task. First, a learning algorithm called FAST for Flexible Adaptable-Size Topology (Pérez-Uribe, 1999) was developed to handle the problem of dynamic categorization of the robots' three 8-bit infra-red "eyes" (which correspond to 24 binary receptors). No external supervisor provides the desired outputs. Second, a trial-and-error learning process coupled with punishment and reward signals (Sutton and Barto, 1998) was considered to allow the robot gen-

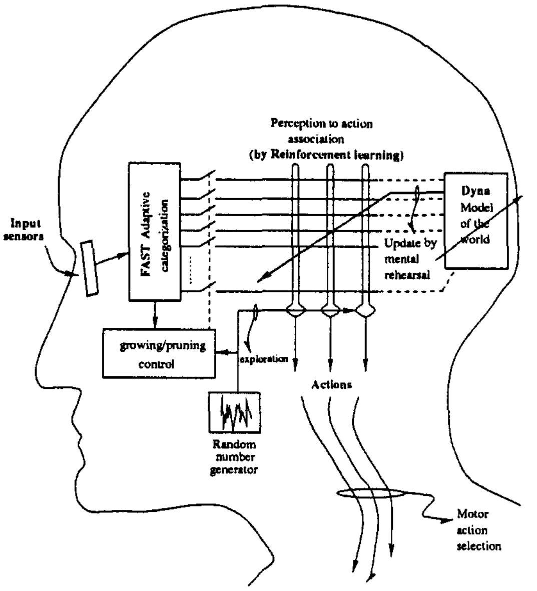


Figure 1: The neurocontroller architecture.

erate behavioral responses as a function of its sensations. Third, a model of the environment is dynamically created to improve the interaction with the actual environment (Sutton and Barto, 1998). The system alternately operates on the environment and on the learned model of the environment by a process of "mental rehearsal".

Finally, we have combined the capabilities of the incremental learning FAST neural architecture with reinforcement learning techniques and planning to learn an obstacle avoidance task with an autonomous mobile robot (Pérez-Uribe and Sanchez, 1999; Pérez-Uribe, 1999).

3 Concluding remarks

We have presented a neural architecture that implements neural epigenesis, reinforcement learning, and mental rehearsal. This architecture may be viewed as a first step towards the development of more complex neurocontrollers implementing many diverse cooperating brain-like structures. Indeed, the implementation of the learning paradigms presented above should enable us to think of a new kind of machines, where, effectively, learning by examples and interaction replace programming (without needing to emulate such principles using a programmable computing machine). In this kind of machines, the learning algorithms would emerge from the dynamics of the interconnection of the processing elements, which may be the key to realize a mind-like system endowed with "semantics" (i.e., a system that is capable of associating a meaning to the symbols it uses for computing) (Searle, 1980, 1990), and not merely with "syntax", as it is the case of our current computing machines.

References

G.E. Edelman. *Bright Air, Brilliant Fire.* Basic Books, New York, 1992.

S. Grossberg. The Link between Brain Learning, Attention, and Conciousness. Technical Report CAS/CNS-TR-97-018, Department of Cognitive and Neural Systems, Boston University, 1998.

D.H. Hubel and T.N. Wiesel. Brain Mechanisms of Vision. *Scientific American*, 241(1), 1979.

R.R. LLinas and D. Pare. Of Dreaming and Wakefulness. *Neuroscience*, 44(3):521–535, 1991.

A. Pérez-Uribe. *Structure-Adaptable Digital Neural Networks.* PhD thesis, Swiss Federal Institute of Technology-Lausanne, EPFL, Lausanne, Switzerland, 1999. Thesis 2052.

A. Pérez-Uribe and E. Sanchez. A Digital Artificial Brain Architecture for Mobile Autonomous Robots. In M. Sugisaka and H. Tanaka, editors, *Proceedings of the Fourth International Symposium on Artificial Life and Robotics AROB'99*, pages 240–243, Oita, Japan, 1999.

S. R. Quartz and T. J. Sejnowski. The neural basis of cognitive development: A constructivism manifesto. *Behavioral and Brain Sciences*, 20(4):537+, 1997.

W. Schultz, P. Dayan, and P. Read Montague. A Neural Substrate of Prediction and Reward. *Science*, 275: 1593–1599, 1997.

J. Searle. Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3:417–424, 1980.

J. Searle. Is the Brain's Mind a Computer Program? *Scientific American*, 262:26–31, 1990.

M. Sipper, E. Sanchez, D. Mange, M. Tomassini, A. Pérez-Uribe, and A. Stauffer. A Phylogenetic, Ontogenetic, and Epigenetic View of Bio-Inspired Hardware Systems. *IEEE Transactions on Evolutionary Computation*, 1(1):83–97, 1997.

R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction.* The MIT Press, 1998.

Functions for the management of valuable goals: Goal scheduling and action mode selection done by an autonomous agent on the basis of situational urgency

Takafumi Tsuchiya
School of Computer and Cognitive Sciences, Chukyo University
101 Tokodachi Kaizu Toyota 470-0393 Japan
tsuchiya@sccs.chukyo-u.ac.jp

Abstract

The purpose of this research is to design an autonomous agent that is required to set and achieve multiple goals with various degrees of urgency in a simple world of a video game type. This poster presentation first introduces the design of our simple world. A task given for the agent is to rescue as many falling objects as possible, which appear randomly in the world. Symbolic descriptions of states constituting the problem space and a plan that the agent would generate, based on the expected utility taking into account the success probability of operation, are given (Feldman and Sproull, 1977). After the formulation of the urgency value for a goal (Toda, 1995), the two phases in the agent's problem solving, the goal scheduling and the action mode selection, are discussed. The goal scheduling produces a quasi-optimal goal queue in a dynamic fashion in accordance with the urgency of the current goal (Minton et al., 1992; Zilberstein, 1996). The action mode selection allocates limited time for actual execution and deliberate planning. High urgency value of the current goal may make the agent stay in the execution mode for a period of the available time. Finally, the current level of implementation and future directions of our research is discussed.

1 Introduction

The functional studies have considered human beings with multiple goals as efficient problem solving systems. While this approach revealed the situated nature of cognitive architecture, it still leaves out some important issues concerning the everyday problem solving. One such issue is that of time, and another is that of the subjective values assigned to achieving each goal, which are necessary for sustaining the unity of an individual human being. In order to deal with the multiple goals, a person ought to be efficient in setting, concentrating, suspending, discarding, and achieving some of possible goals in accordance with the person's cognitive appraisal of the urgency with which each goal presents itself. The urgency of a goal is an important situational cognition made by a problem solver with a limited temporal resource. During the period when no action is taken, the urgency of each goal will increase as the available time for accomplishing the goal-achievement action diminishes.

The purpose of this research is to design an autonomous agent that is required to set and achieve multiple goals with various degrees of urgency in a simple world of a video game type. For the functional study of architecture, Simon (1967) discussed an interruption mechanism of ongoing processes on a serially fashioned cognitive architecture. This research employs a serially fashioned architecture for coping with situations in the simplified world, and intends to specify various functions for the management of goals.

The agent embedded in the world is designed to have three phases in its course of problem solving. The first phase is planning to make a better plan searched as a solution path of operators in the problem space for achieving each single goal. The second phase is goal scheduling in the face of multiple goals, whose function is to schedule how to achieve the given set of goals in what order. Note that, while the target goal is being achieved, the urgency values of other goals in queue will increase due to the decrement in the available time for their achievements. The scheduling rule by a heuristics called urgency comparison is proposed. What it is aimed to do is to reduce the sum of urgency values of all the goals. The third phase, that of action mode selection, does the switching of its action mode between the execution mode and the deliberation mode to be done in accordance with the urgency presented by the current goal. If this urgency value is very high, the agent should allocate its time for rush execution of some operators in a plan, despite of its limited plausibility. On the other hand, if the urgency is relatively low, the agent may be able to engage in a more deliberate appraisal of the global situation.

The poster presentation deals mainly the two phrases in the problem solving of the agent, namely, the goal scheduling phase and the action selection phase, with the example of one such agent implemented in a visualized world. In the following, the design of our world and the formulation of the urgency value for a goal are briefly introduced.

2 The world and urgency

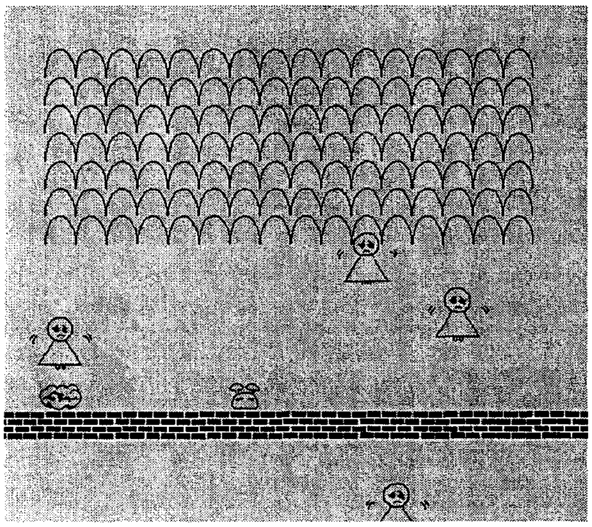


Figure1: The snapshot of the world

Figure 1 shows a snapshot of this world. An agent appears around the center on the floor. A task given for the agent is to rescue as many falling objects as possible. There are four falling objects in Figure 1, one of which has already passed the floor and cannot be rescued. The agent can move horizontally on the floor and put under a falling object a life preserver looking like a cloud in Figure 1. The success probability of this operation called "PUT-LIFE-PRESERVER" is 0.5. A life preserver can be put only if an object is falling over it. A preserver disappears, once it is used for an object. A new object appears randomly. While an object is falling, no other object appears in the same row. The speeds of the agent's moving and of the object's falling are fixed.

If a goal is assigned to rescuing a particular object, the planner in the agent would generate a plan, which is based on the expected utility taking into account the success probability of PUT-LIFE-PRESERVER operation (Feldman and Sproull, 1977), a plan such as "repeat MOVE operation of going under the object and then repeat PUT-LIFE-PRESERVER operation until it succeeds."

The urgency of each goal at a given time is defined by the following three parameters, namely, the subjective value of each goal, the subjective probability of achieving it, and the available time for doing so. By an extension of the classical decision making formula to make it include time, Toda (1995) defined expectation of the subjective value of a goal at available time t:

$$V_t = P_t V,$$

where

V is the value of a goal when it is actually gained.

P_t is success probability to achieve that goal at available time t; $0 \le P_t \le 1$.

Urgency value of a goal at available time t is defined as the expectation of the subjective value that could be lost:

$$U_t = (1 - P_t)V,$$

The value of P_t is defined in terms of the two parameters, the subjective probability for the success of prescribed operation and the available time. The formula is

$$P_t = P_t(\mathrm{MOVE})P_t(\mathrm{PUT-LIFE-PRESERVER}),$$

where

$P_t(\mathrm{MOVE})$ is success probability of MOVE operation given the available time.

$P_t(\mathrm{PUT-LIFE-PRESERVER})$ is success probability of PUT-LIFE-PRESERVER operation given the available time.

If the agent has time for a MOVE operation, the available time for PUT-LIFE-PRESERVE operation is determined by the following three temporal parameters: the time that a falling object takes to reach the floor, the required time for MOVE operation, and the time needed for information processing done by the agent. If the agent spares the available time to repeat PUT-LIFE-PRESERVE operation until it succeeds the urgency value of the goal decreases and the expected goal value increases.

References

J. A. Feldman and R. F. Sproull. Decision theory and artificial intelligence II: the hungry monkey. *Cognitive Science*, 1:58-192, 1977.

S. Minton, M. D. Johnston, A. B. Philips and P. Laird. Minimizing conflicts : a heuristic repair method for constraint satisfaction and scheduling problems. *Artificial Intelligence*, 58:161-205, 1992.

H. A. Simon. Motivational and emotional controls of cognition. *Psychological Review*, 74:29-39, 1967.

M. Toda. A decision theoretical model of urge operations. from Toyota as Chukyo University SCCS Tech. Rep. No.95-1-01, 1995.

S. Zilberstein. Using anytime algorithms in intelligent systems. *AI magazine*, 17:73-83, 1996.