Integrative Approaches to Machine Consciousness

5th - 6th April 2006

Organisers

Rob Clowes, University of Sussex
Ron Chrisley, University of Sussex
Steve Torrance, Middlesex University

Programme Committee

Igor Aleksander, Imperial College Lond.
Giovanna Colombetti, York University
Rodney Cotterill, Technical University of Denmark
Frédéric Kaplan, Sony Computer Science Laboratory
Pentti Haikonen, Nokia Research Center
Germund Hesslow, Lund University
Owen Holland, University of Essex
Takashi Ikegami, University of Tokyo
Miguel Salichs, University Carlos III
Ricardo Sanz, Polytechnic University of Madrid
Murray Shanahan, Imperial College London
Jun Tani, Brain Science Institute
Steve Torrance, University of Middlesex
Tom Ziemke, University of Skövde

Contents

On Architectures for Synthetic Phenomenology

Igor Aleksander
Dept. Of Electrical and Electronic Engineering,
Imperial College , London SW7 2BT
i.aleksander@imperial.ac.uk

Helen Morton
School of Social Sciences and Law
Brunel University, Uxbridge UB83PH
Also, Imperial College , London SW7 2BT
helen.morton@brunel.ac.uk

Abstract

Is synthetic phenomenology a valid concept? In approaching consciousness from a computational point of view, the question of phenomenology is not often explicitly addressed. In this paper we review the use of phenomenology as a philosophical and a cognitive construct in order to have a meaningful transfer of the concept into the computational domain. Two architectures are discussed with respect to these definitions: our 'kernel, axiomatic' structure and the widely quoted 'Global Workspace' scheme. The conclusion suggests that architectures with phenomenal properties genuinely address the issue of modelling consciousness and indicate and the way that a machine with synthetic phenomenology may benefit from the property

1 Introduction

In searching for computational models of being conscious, the detailed nature of internal representation is an important facet of the way that modelling is to be approached. Synthetic phenomenology is involved when two conditions are fulfilled: first there is a meaningful sense in which a first person may be ascribed to the model and second, when the architecture caters for an explicitable and action-usable representation of "the way things seem" within the machine. We take the view that rather than this being an idealist stance, it represents as close an approximation to "the way things are" as is permitted by the sensory apparatus of that organism. This is assumed to be sufficiently close to reality to enable the organism to take appropriate action in its world. So one expects to find accurate phenomenological representation in successfully evolved organisms, as a major distance between the representation and reality does not augur well for successful evolution.

The paper first reviews the reason that in philosophy, phenomenology had a firm foothold despite the fact that the appellative became used in a variety of ways. A brief discussion is included on Block's use of the word in the notion of 'Phenomenal consciousness' as being distinct from 'Access consciousness' and, particularly in the way that such concepts could feature in computational systems.

The concept of a 'depictive' representation is developed in this paper beyond that which has been discussed to date (Aleksander, 2005) to show that this is a central requirement for an architecture that could be said to be synthetically phenomenological. A set of architectural definitions is then developed that determines whether an architecture could be said to be phenomenological or not. Two known architectures are scrutinised from the point of view of these definitions: are own *kernel* architecture (Aleksander, 2005) and Shanahan's embodied version of Baars' Global Workspace architecture (Shanahan, 2005). This reveals that the issue of phenomenology can be considered for differing mechanistic descriptions, of which the two architectures are distinct examples. In the conclusion we argue that the material in the paper indicates that architectures that are phenomenological have characteristics of being conscious that enhance their use both as explanatory tools and, possibly, functional artefacts. We shall first review issues that go under the heading of Phenomenology *and italicise strands that are*

taken up in discussing the implication for synthetic systems and their architectures discussed later in the paper.

2 Phenomenology

2.1 Definition

In the broadest terms, phenomenology is the word given to studies of consciousness which specifically start with the first person. In other words, introspection is an important facet of the discussion. This distinguishes phenomenology from other forms of philosophy, say, ontology, which asks what it is for an object to *be* conscious. One should also distinguish 'a phenomenon' from other philosophical constructs such as 'qualia' which relate to sensational primitives such as 'redness' or 'the sweet smell of a rose'. In general, phenomenologists like to extend the definition beyond the immediate sensation to more compositional structures of experience such as enjoying a game of tennis or the experience of having tried a new restaurant. This also aids action in the world and the generation of descriptive language in the case of humans or human-like machines.

Conforming with the above definition, the 'kernel' architecture we shall discuss in this paper was synthesised through a process of using introspection to discover design principles. This led to a consideration of ways that this work contributes to the formation of a synthetic phenomenology paradigm.

2.2 Past Usage

It is noted that in the history of philosophy, phenomenology is sometimes treated as the study of consciousness itself. For Franz Brentano (1874 trans. 1995) phenomena *are* acts of consciousness, they are the contents of mind. They stand in relation to physical phenomena that are perceived in the world by intentionally creating meaning of physical elements of the world in the mind. This first-person, descriptive character of a phenomenon has remained the hallmark of the work of later phenomenologists. Of these, Edmund Husserl (1913 trans. 1989), also focuses on the meanings the mind creates when contemplating the real world. This position addresses the mental object beyond just its real-world shape. So a stick may have the ability to dislodge a banana off the branch of a tree, enhancing the phenomenology of the stick by a mental vignette of the action of dislodging the banana.

Martin Heidegger (1975, trans. 1982) maintained that setting ontology (what it is to be conscious) apart from phenomenology could be an error. He suggests that it is actually linked to the phenomenology of the first person sensation of being a self in an external world. *See the influence of this in what we shall call 'axiom 1'.* Given Sartre's socio-philosophical observations on phenomenology as a literary examination of one's own experience and Maurice Merlau-Ponty's linking of phenomenology to personal experiences of one's own body (1945, trans. 1996) this becomes important particularly for those who discuss consciousness in the context of embodied robots.

The body's muscular activity is a key element in the 'kernel' architecture to create 'depictions', that is sensations of being an entity in an out-there world. As will be seen, Shanahan argues that embodiment is essential to have an experiencer.

2.3 Materialist Concerns

Gilbert Ryle in *Concept of Mind* (1949) argued that linguistic descriptions of mental states are a direct way of expressing phenomenology. This was possibly erroneously discredited by many materialists who identified the mental state with the neural state. Clearly only some neural states support phenomenology as identified by Crick and Koch (2003). Only some parts of the entire neural state are responsible for personal sensation, the parts that are not, have been called by the authors the 'Zombie' regions of the brain. This appears to beg the question of how one distinguishes a neuron that contributes to conscious sensation from one that does not. *A possible answer was developed by Aleksander and Dunmall (2003) and Aleksander (2005). This draws attention to the fact that in the visual system only some neurons, those indexed by the motor areas of the brain, can fire in a way that correlates with elements of the visual sensation of being an entity in an 'out-there' world. This is summarised later in this paper.*

2.4 Access and Phenomenal Aspects

Ned Block (1995) has identified at least two salient functions of consciousness. The first he calls 'phenomenal' or P-consciousness to indicate the personal function of experiencing a mental state. He contrasts this with 'Access' or A-consciousness which is that function of consciousness which is available for use in reasoning, being 'poised' for action and the generation of language. Although he argues that both are present most of the time, conflating the two when studying consciousness is a severe error. Some evidence of Block's distinct forms of is drawn from the phenomenon of 'blindsight' where individuals with a damaged primary visual cortex can respond to input without reporting an experience of the input.

This is A without P. P without A is the effect that some unattended experience had happened previously (e.g. a clock striking) but the individual had only realised this later. That is, P without A covers the case that unattended input can be retrieved. This creates a mechanistic difficulty for the definition of phenomenal consciousness as, were it never to be brought into access format, it could not in any way be described as 'the way things seem'. In hard-nosed synthetic phenomenology it might be politic to concentrate only on things that have seemed or seem to be like something.

This implies that in architectures it is important to be clear about the way in which immediate perceptual consciousness interacts with awareness of past experience, which bears on the A/P discussion.

Blindsight has also entered the theories of 'enacted' vision proposed by Kevin O'Regan and Alva Noë (2001) who have broadly argued that 'representing' the visual world in any architecture, living or synthetic, is an error, as the world itself is representation enough for the system to act on in a physical way. Consciousness is then a 'breaking into' this somewhat reactive, autonomic process through mechanisms of attention.

It is known that in the brain there are unconscious sensorimotor processes of the O'Regan and Noë description that work in conjunction with conscious phenomenal processes. For example the oculo-motor loop that involves the superior colliculus is such a mechanism. We are not conscious of the retinal maps that are projected onto the superior colliculus. They lead, also unconsciously, to the saliency maps that partly determine eye movement which eventually leads to reconstructions of world-fixed representations much deeper in the visual cortex (the extrastriate regions according to Crck and Koch, 2003). The enacted-unconscious/depicted-conscious interaction is a useful concept that may be used in synthetic systems. We find it difficult to accept the 'hard' sensorimotor view that complete access to a visual world can be achieved without any phenomenal representation at all.

3. Phenomenology in Computational Models

There are two important computational issues we wish to stress here. The first is the nature of a third-person design of an object that is capable of first-person representation, and the second is the relationship of depiction to synthetic phenolenology.

3.1 The Third Person Design with First Person Within It.

Where, in philosophy, phenomenology starts with the first person sensation, we suggest that in computational modelling, a phenomenological model must, in the broadest terms, sustain representations that have first person properties for *the model itself*. There is no dualist slight of hand here as the designer of the system can happily retain a third-person view of what is being designed, given a theory of what in the design is necessary to achieve a first person for the mechanism. That is, despite starting with our own first-person sense, we can speak of the first person of others. Similarly, we can speak of the first person of a machine and, indeed, set out to search for mechanisms of such. This implies that, in vision, for example, there is a need to differentiate mechanisms that mediate the sense of presence of the organism in the world from those that are due to previous experience: memory of various kinds and imagination (for example, states induced by literature). That is, there needs to be computational clarity about how a first-person phenomenal state relates to the current world event, how meaning is assigned to this, how meaningful states arise even in the absence of meaningful sensory input and how a personal sensation of decisions about 'what to do next' can arise. In Aleksander and Dunmall, 2003 and Aleksander 2005 we have referred to a necessary property for the machine having a first person at all as being a 'depiction'. Here we set out this concept as a logical sequence.

3.2 Depiction and Phenomenology.

It is useful to define what we mean by a *synthetically phenomenological system*.

> **Def 1**: To be **synthetically phenomenological**, a system S must contain machinery that represents what the world and the system S within it *seem* like, from the point of view of S.

The word *seem* has been transferred from the phraseology of the earlier parts of this paper to stress that perfect knowledge of the world cannot be achieved if only because of the weaknesses of sensory transducers. But, it is stressed that living creatures, if we believe that they have phenomenological representations, will come to our notice only through successful evolution. Again we stress that this is due to some sufficiency in the similarity between what things seem like and how, in a sense important to the organism, they not only *seem like* but, as far as the organism is concerned, *they are.* To achieve this it is necessary that such a representation should fully compensate for trans-

ducer and body mobility. In earlier work we have called this a 'depiction' rather than a representation. To advance this prior work we develop a series of definitions and assertions about depictions that positions this work within the framework of phenomenology addressed earlier.

Def 2: A **depiction** is a state in system S that represents, as accurately as required by the purposes of S, the world from a virtual point of view within S.

Assertion 1: A depiction of Def. 2 defines the mechanism that is necessary to satisfy that a system be synthetically phenomenological according to Def. 1.

Assertion 2: If S is mobile and has mobile sensors, a depiction of Def. 2 can only be achieved if the mobile nature of S is combined with the information carried by the sensors. That is the 'where' of the elements of the world needs to be predicated on the 'body' parameters of S. (In vision, eye-movement clearly needs to be compensated to achieve a depiction).

Assertion 3: 'As accurately as required ..' in Def. 2, indicates that, given effectors with which to act on the world, the depiction should carry all the information needed for such effectors to be successfully deployed on the attended and desired elements of the world.

Assertion 4: 'As accurately as required ..' also sets determines the granularity with which the depiction may be achieved.

Assertion 5: While Def. 2 makes no call on a topological representation, it does require that differently positioned elements within the representation be indexed by the predicates introduced in assertion 2. In animal vision it is known that different attributes of a visual element (e.g. the colour and motion of a dot) are represented in different parts of the brain. What 'binds' them in our analysis is the indexing as clarified in the example below (see Aleksander and Dunmall, 2000).

Example of indexing: Participant X is fixating a cross in the centre of a screen. She is asked to identify the shape *s* and colour *c* of an object that will appear briefly on some other part of the monitor screen. Shape is represented in area P of her brain and colour in area Q. The eye driven by the superior colliculus will saccade to the position of the object. The signal issued by the eye movement is, say, a 2-dimentional vector $\boldsymbol{v}$. Then the depiction in P will be *s,* indexed by $\boldsymbol{v}$, say s_v . Similarly, in Q we have c_v . Assertion 5 states that the binding of *s* and *c* is due to the common indexing by $\boldsymbol{v}$: that is, $(s,c)_v$.

It is the deeper contention of the depictive approach that $(s,c)_v$ uniquely encodes X's phenomenal experience of the appeared object. Of course, away from this experimental example, the indexing, as indicated by a great deal of physiological evidence (e.g. Galletti & Battaglini, 1989) occurs over many areas of the cortex, giving the phenomenal experience of one sensory modality several dimensions possibly bound across modality boudaries. Touch together with vision are a commonly bound experience.

4. Architectures

By 'architecture' we refer to a structure that first, is made of several internal parts each of which performs a specified distinct function, and second, includes a full specification of the interconnections among these parts the inputs and a variety of outputs (e.g. language generators, physical actuators etc..). It is the contention of this paper that there exists a set of architectures that can support phenomenology for the organism that embodies the architecture. We shall first look at two specific architectures to assess some of the definitional material presented in section 3.

4.1 The 'Kernel' Architecture

It is hardly a coincidence that a prototypical architecture we have recently suggested (Aleksander, 2005) should be based on the notion of a depiction and can, therefore, be said to have phenomenal consciousness according to our criteria. We take a closer look at this scheme that is shown in Fig.1

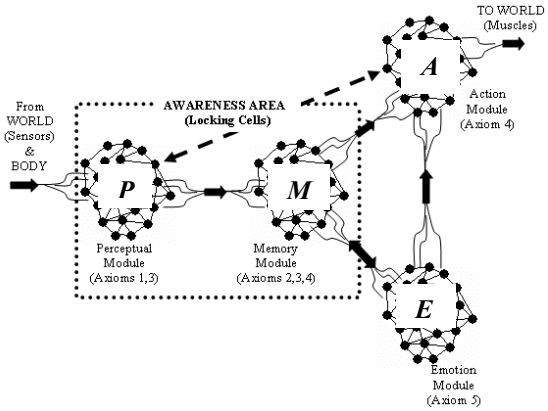

Figure 1. The 'kernel' architecture.

This architecture is based on the axioms of consciousness published in Aleksander and Dunmall (2003). For completeness, they are briefly listed in the Appendix of this paper.

These axioms start from a phenomenological standpoint as they are derived through an intro-

spective decomposition of the most significantly felt aspects of being conscious. Then it has been argued that the decomposition eases the transfer of these features into the synthetic domain.

Fig.1 is the result of this process. It consists of five modules each of which is considered to be a neural state machine (NSM) that operates in binary mode. That is, each connection carries a binary signal. We have ofen argued that any loss of generality due to the binary synthesis will be minor with respect to the behaviours that are being researched.

The binary NSM is specified as a six-tuple:
$<Ci,Co,Cf,Ct, I, O, F, T>_n$ where,
n is the module index,
Ci is a connection pattern of inputs (which may come from other modules or sensory inputs);
Co is a connection pattern of outputs (to other modules or system outputs);
Cf is the pattern of internal feedback connections.
Ct is the set of 'teaching connections' that determine the state of Co and Cf that becomes associated with Ci.
I, O, F, and T are the state sets of Ci,Co,Cf and Ct respectively.

Then, in the usual way with neural state machines, the states of F(t) and O(t) become functions of F (t-1) and I(t). These functions are determined by a training strategy which is expressed through T during a 'training phase'.

For example, an 'Iconic' mode of training is conventional with neural state machines of this kind (Aleksander and Morton, 1995). This ensures that, given that Ct and Cf have the same dimensions and Co=Cf, the network learns F(t)=T(t) as a function of I(t) and F(t-1).

Returning to Fig.1, the four axioms are implemented as follows. P is a 'Perceptual' NSM which is made to be phenomenological in the sense of the earlier definitions of this paper through the following design. The state F(t) is a reconstruction of the sequences of attended world inputs from sensory transducers over defined time windows (sometimes sliding time windows). The muscular effort required to attend to the elements of the world is shown as the link from the action NSM, A. In the animal visual system it is surmised that attentional shifts are driven by saliency maps in the superior colliculus. In specific studies of the visual system, this has been modelled as an additional part of the kernel architecture (See Igor Aleksander et al. 2001)

M is the memory and 'imagination' module. It is connected to P in such a way that for every reconstruction in P, a state in M is created. Sequences of reconstructed states in P can therefore be stored as state trajectories in M – they will have inherited the depictive, hence phenomenal properties of P.

P and M together form what we have dubbed 'the awareness areas' of the architecture. In the sense that one can perceive and recall at the same time, the two areas both contribute to the same phenomenal state The remaining modules of the kernel architecture are not depictive, hence not phenomenal, but add to the phenomenal existence of the system in the following way. As mentioned, A is the action area in which links between the state trajectories of the phenomenal areas are translated into action. But this is not automatic, it is surmised that volition and emotion as implemented in module E mediate this link. This was the subject of the contribution by Aleksander, Lahnstein and Lee in the AISB 2005 symposium on machine consciousness.

In summary, the kernel architecture is based *ab initio* on the intention of synthesising an architecture with phenomenological properties. This has also been guided by those who like Crick and Koch (2003) have been researching the neural correlates of consciousness in living organisms. We now consider a model that is more closely related to computational approaches of the functional kind.

4.2 Embodied Global Workspace

Bernard Baars' (1988, 1997) Global Workspace models have held sway in computational modelling of consciousness for some years. Baars considered how a large number of unconscious processes might collaborate to produce a continuum of conscious experience. In very broad terms, he answers the question through the architecture of Fig. 2.

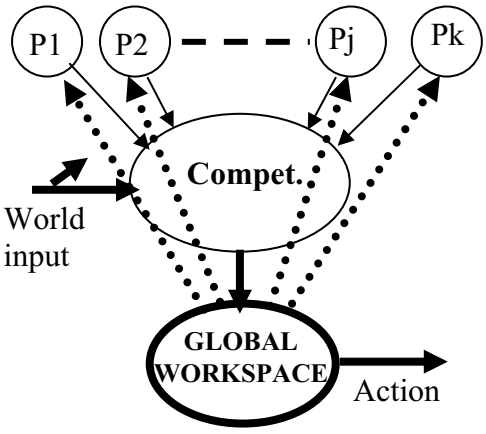

Figure 2 A sketch of Baars' Global workspace architecture.

The separate processes, P1 to Pk, said to be unconscious, compete to enter 'The Global Workspace'.

Such processes are often thought of as memory activities, say, episodic memory, working memory and so on. The competition is won by the process that has the greatest saliency at a given moment. This saliency is predicated by world input which sets the context for the competition. Of course, world input is also assumed to have direct influence on the unconscious processes P1 - Pk. Having entered the global workspace, the winner of the competition becomes the conscious state of the system. This is continuously 'broadacast' back to the originating processes that change their state according to the conscious state. This results in a new conscious state and so on, linking sensory input to memory and the conscious state. It is both general and useful for these separate processes to be modelled as NSMs as was done for the kernel architectures

Murray Shanahan (2005) points out that modelling of a conscious organism cannot proceed without that organism being embodied in some palpable world. Using the above Global Workspace model he argues that there can be no 'experiencer' in GW unless the model takes account of the "spatial unity of the body". It is this localisation in space that for Shanahan gives the model its "viewpoint on the world" which according to def. 1 makes it a candidate for phenomenal consciousness. Shanahan argues that denying this possibility, as is done by Block (1995), revives the dualist stance, putting phenomenal consciousness in the Chalmers-like 'hard problem' class, that is, a problem that cannot be reduced to physical structure and hence cannot be synthesized. And yet, the claimed 'point of view' of the embodied organism is undoubtedly a claim that this accords with definition 1 above of a phenomenal system. In terms Block's division into access and phenomenal consciousness, Shanahan implies that the embodied GW model addresses access consciousness, treating the phenomenal element as being an unnecessary appeal to a dualistic concept.

4.3 GW and Synthetic Phenomenology

While it seems entirely correct that without embodiment, GW does not include an experiencer, the question remains of how the experience stream in GW relates to the real world. We recall that in section 3.2 we have argued that a synthetic phenomenological system is achieved through a compositional representation of the world that is sufficiently accurate for the system be able to use its embodiment to control its world as accurately as possible. That is, it is the contention of this paper that depiction is the missing ingredient in making GW phenomenal. That is, phenomenal consciousness can occur in functional, physical systems, and the implication for the embodied GW system is that ***all*** the P1-Pk states need to be ***depictive*** for the GW state to be truly a model of a conscious state. Were this not the case, some translation into depiction would have to go along with the winning of the competition. Otherwise the spectre of purely arbitrary representations in GW remains. Shanahan is aware of this by requiring that the conscious broadcast back to the competing processes be in some way intelligible to these processes. But this still makes it hard to see how the states of the processes remain non-depictive when the state of GW might be depictive.

5. Discussion

In this paper we have explored the concept of synthetic phenomenology mainly by attempting to define the necessary features of an architecture that supports phenomenal consciousness within the broadest definition of the term. We brought the definitions to ground by considering two models that might be candidates for possessing these features. To conclude we raise and, using the material of this paper, attempt to answer five general questions that may be central to the existence of a synthetic phenomenology. The first of these addresses the architectures presented in the paper.

Can non-depictive representations be phenomenal?

It is the firm implication of this paper that this cannot be the case. It is depiction in a functional area which determines that the area contributes to the phenomenal sensation of the organism. Were this not the case, a human description of a state would require translation into phenomenal terms as such descriptions are of phenomena and not encoded states.

What is the difference between 'depictive kernel' and GW architectures in terms of synthetic phenomenology?

Clearly the depictive kernel architecture was designed with the purpose of creating a phenomenal representation within the system according to the definitions set out in this paper. This has the computational advantage of being able to be display on a screen the current phenomenal state of the machine enabling a designer's assessment of the interactions between both postulated conscious and postulated unconscious mechanisms in the generation of the phenomenology. The rules used in the synthesis involve depiction. Originally no phenomenal claims were made for GW, particularly in its practical form as synthesised by Stan Franklin (2003). However with the embodied GW work of Murray Shanahan, the question of the presence synthetic phenomenal consciousness acquires a

new urgency. In this paper we have maintained that were an architecture based on GW to have a phenomenal character, there must be a depictive activity in the processes that compete for entering the global workspace if the system is to be phenomenological. This creates problems as in our scheme of things, depiction in an area of the architecture implies phenomenal consciousness and GW sees the competing processes as being non-conscious. Therefore a phenomenal GW implies some sort of coming into consciousness in the GW area for reasons other than depiction. These have not yet been explained. Of course, the depiction idea can be rejected, but if not depiction, then what?

What is the use of synthetic phenomenology? Given the difficulties mentioned with embodied GW above, it is proper to ask why bother with phenomenology and why not settle for just access consciousness as implied by Shanahan (2005)? In the arguments of the current paper, phenomenology actually includes the purposes that are attributed to access consciousness. But such purposes are explicit and searchable through attentional mechanisms for reasons of accurate interaction with the environment (see assertions 3 and 4). This is not a Blockian confusion, but rather a suggestion that there may not be as clear-cut a functional/neurological distinction between access and phenomenal consciousness as Block seems to suggest. The A without P and P without A cases may be extreme conditions of a central phenomenon. In summary we argue that accurate interaction with, and thought about the real world is the purpose of phenomenology in a synthetic system.

Is synthetic phenomenology an oxymoron as it is the non-physical experiential side of consciousness and therefore eschews synthesis? Everything we have submitted in this article is a denial of the above proposition. Treating phenomenology as the 'hard' part of consciousness simply kicks it out of touch of science into some mystical outfield. We maintain that addressing it as a constructible concept removes the mysticism with which it might otherwise be associated.

Is synthetic phenomenology an arbitrary design option for models of consciousness? This paper regards models of consciousness without synthetic phenomenology as being valid only in a behavioural sense. That is, it is possible for a model to be given attributes of being conscious from its behaviour. Stan Franklin's Intelligent Distribution Agent (2003)is a good example of this class of system. Users think that they are dealing with an entity *conscious* of their needs. But if one were to argue that an architecture throws light on the mechanisms of consciousness in the brain it becomes mandatory to include phenomenal, that is depictive functions.

What research needs to be done in developing architectures with synthetic phenomenology? Referring to the kernel architecture there is much work to be done on modes of interaction between the modules. Current work includes a clarification of the way the emotion module E controls the link between the phenomenological P and M modules and the non-phenomenological action module, A. (fig. 1).

Illusions, ambiguous and 'flipping' figures are situations where phenomenology and reality part company. We are pursuing the mechanisms that, in the kernel architecture, would lead to the kind of perceptual instabilities associated with perceiving the Necker cube. This underlines the usefulness of synthetic phenomenology, as perceptual reversals may be measured in the depictive machinery and the conditions for such reversals studied. This is revealing of the interaction between phenomenal and non-phenomenal processes in the brain

In GW, architectures it would be interesting to clarify the causes of phenomenology in the GW area which are not present in the supporting competitive processes.

Appendix: Axioms of Being Conscious.

This is an introspective partitioning of five important aspects of being conscious

1. I feel as if I am at the focus of an out-there world.
2. I can recall and imagine experiences of feeling in an out there world.
3. My experiences in 2 are dictated by attention and attention is involved in recall.
4. I can imagine several ways of acting in the future.
5. I can evaluate emotionally ways of acting into the future in order to act in some purposive way.

References

Igor Aleksander, *The World In My Mind, My Mind In The World*' Exeter: Imprint Academic, 2005.

Igor Aleksander, Mecedes Lahnstein, Rabinder Lee: Will and Emotions: A Machine Model that Shuns Illusions, Proc AISB 2005 Symposium on New Generation Approaches to Machine Consciousness, 2005

Igor Aleksander, and Barry Dunmall: Axioms and Tests for the Presence of Minimal Con-

sciousness in Agents *Journal of Consciousness Studies.* **10**, pp 7-18, 2003

Igor Aleksander, Helen Morton and Barry Dunmall Seeing is Believing. *Proc. IWANN01*, Springer, 2001

Igor Aleksander, and Barry Dunmall:). An extension to the Hypothesis of the Asynchrony of Visual Consciousness, *Proceedings of the Royal Society ofLondon B* **267**: 200, 197–200.

Igor Aleksander and Helen Morton, *Introduction to Neural Computing (2nd Edition)*, London: Thomson Computer Press, 1995

Bernard Baars, In the Theater of Consciousness: The Workspace of the Mind , New York: Oxford University Press, 1997.

Bernard Baars, *A Cognitive Theory of Consiousness* , Cambridge: Cambridge University Press, 1988.

Ned Block, On a Confusion about a function of Consiousness, *Behavioural and Brain Sciences,* **18**, pp 227-287, 1995

Franz Brentano, *Psychology from an Emptirical Standpoint*, Trans: Rancurello et al. Routledge, 1995, Orig in German 1874.

Francis Crick and Christof Koch, 'A Framework For Consciousness' *Nature Neuroscience ,***6**, pp119 – 126, 2003 .

Stan Franklin, 'IDA a Conscious Artifact?' *Journal of Consciousness Studies,***10** (4-5), pp47-66, (2003)

Claudio Galletti and Paolo Battaglini: Gaze-Dependent Visual Neurons in Area V3A of Monkey Prestriate Cortex. *Journal of Neuroscience*, **6**, 1112-1125, 1989

Martin Heidegger, *The Basic Problems of Phenomenology,* Trans Hofstadter, Indiana University Press, Orig in German, 1975.

Edmund Husserl, *Ideas:A General Introduction to Pure Phenomenology,* Trans. Boyce Gibson, Collier, 1963. Orig in German, 1913.

Maurice Merlau-Ponty, *Phenomenology of Perception,* Trans Smith, Rotledge 1996, Orig in French, 1945.

Kevin O'Regan and Alva Noë, ., A Sensorimotor account of vision and visual consciousness. *Brain and Behavioural Sciences***, 24**(5) 2001.

Gilbert Ryle, *A Concept of Mind,* London: Hutchinson's, 1949.

Murray Shanahan, 'Global Access, Embodiment and the Conscious Subject'. *Jour. Of Consciousness Studies*, **12**, No 12, 2005 (in press)

Correlation, Explanation and Consciousness

Margaret Boden
Centre for Research in Cognitive Science
University of Sussex,
Falmer, Brighton, Sussex BN1 9QH, UK
maggieb@sussex.ac.uk

. **Abstract**

There's a lot of excitement about brain-scanning evidence for brain/consciousness correlations. Although the evidence is new, the idea isn't: Descartes formulated it nearly 400 years ago. However, he didn't regard mind-brain correlations as explanations – and neither should we.

Mere correlation between events in two domains is not enough for the one to be used as an explanation of the other. In addition, we need systematicity, isomorphism, and plausible (ideally, predictive) counterfactual conditionals.

There are a few (very few) examples where we already have those features, in respect of correlations between brain events and consciousness. In general, however, they can't be expected.

Even where we do have them, they leave the most difficult problem about conscious experience untouched.

The Problem of Inner Speech and its relation to the Organization of Conscious Experience: a Self-Regulation Model.

Robert Clowes
Centre for Research in Cognitive Science
Department of Informatics
Sussex University
Brighton BN1 9QH
East Sussex
UK
robertc@sussex.ac.uk

Abstract

This paper argues for the importance of inner speech in a proper understanding of the structure of human conscious experience. It reviews one recent attempt to build a model of inner speech based on a grammaticisation (Steels, 2003). The Steels model is compared with a *self-regulation* model here proposed. This latter model is located within the broader literature on consciousness. I argue the role of language in consciousness is not limited to checking the grammatical correctness of prospective utterances, before they are spoken. Rather, it is more broadly activity structuring, regulating and shaping the ongoing structure of human activity in the world. Through linking inner speech to the control of attention, I argue the study of the functional role of inner speech should be a central area of analysis in our attempt to understand the development and qualitative character of human consciousness.

1 Introduction

To introspection, for many of us, our mental life seems to have a constant accompaniment of inner speech. This speech is known in the literature under a number of names such as; the inner voice, the internal monologue, and is sometimes, subsumed into (the more general) stream of consciousness (James, 1890). It may also be linked to the generally pejoratively associate notion of 'voices in the head'. Understanding the nature of this phenomenon and its functional underpinnings, although of occasional interest in the history of psychology, has, in the last few years drawn the attention of many researchers into mind. There is however, much controversy about the precise nature of inner speech, its epistemic status and possible functional role.

Among psychologists, one means of accounting for inner speech is Baddeley's articulatory loop (Baddeley & Hitch, 1974), later rechristened the phonological loop[1] (Baddeley, 1997). This is considered to be a speech related working memory system.

Among philosophers, the notion of inner speech suggests privileged access to mental states, and this, at least in the 20th century, has invited great scepticism. The high-water marks of this scepticism are probably Ryle's (1949) *The Concept of Mind* and Dennett's (1991) *Consciousness Explained.* Dennett's view is complex on this question for although he ultimately doubts the strength of the epistemic warrant that can be given to the narrative stream of consciousness, and especially the subject's privileged position to report on its contents, he nevertheless argues that the subject's self-reports should be our starting-point. This is fundamental to his *heterophenomenological* method. This approach advocates

[1] Presumably this re-naming has something to do with thinking of inner speech as primarily an imaged sound, rather than unvoiced speech. The notion of a phonological loop seems to focus on the phenomenology of the passive, rather than active aspect of inner speech.

we need to attempt to offer some explanation of the importance attached to inner speech in phenomenological accounts.

A window into the phenomenology of inner speech is provided by Russell Hurlburt's *Descriptive Experience Sampling* technique (1990). Hurlburt uses an experimental technique in which subjects are cued by a small alarm device at various moments in their day, and then following protocols developed by Hurlburt, write down the details of their mental imagery at the moment that the alarm went off. He argues this technique allows us to systematically sample the qualitative characteristics of reported phenomenology[2]. It also allows us to describe some of the characteristics of inner speech, and inner imagery in general, in a much more elaborated fashion.

The content and form of this reported inner speech seems to be very diverse. Some people report the perception of being the author of voice-like inner speech; others, to hearing voices offering advice or consolation. Sometimes this voice appears to be their own, and sometimes the voice of another person. Some people report merely having the sense of experiencing language-like cognitive episodes without necessarily hearing any voices or having the sense of being the author of this speech. The variety of this speech might serve as some justification for the sceptics, or perhaps just evidence of the complexity and variety of the roles played by speech in our mental lives.

All of these phenomena seem to vary considerably both across individuals, within individuals at different times and places, and with regard to whatever activities they are at that moment engaged in. Hurlburt's work reveals much of the contents of consciousness appear to be composed of speech-like episodes. Except in cases of severe psychological disturbance or other abnormal functioning, the inner voice seems to be the constant accompaniment of human conscious life. But can we relate these accounts of the contents of conscious experience to language as vehicle?

Some recent accounts of cognitive role of language have brought to the fore they way that language may play a role, in sculpting, stabilising, and supporting forms of thought which would be otherwise impossible (Carruthers, 2002; Clark, 2004). Trying to forge a link theoretically between the phenomenological and functional aspects of inner-speech has proved so far a difficult task, but it is one upon which some progress has now started to be made.

2 – A re-entrance model of inner speech

Although traditional work on cognitive modelling made much use of more-or-less linguaform internal representations, following (if sometimes implicitly) some version of Fodor's (1975) Language Of Thought hypothesis, it has shied away from explicitly modelling the inner voice (cf. Dennett, 1994). Perhaps this is because of a worry that the inner voice might be either an epiphenomenon or user "illusion" (Dennett, 1991).

Recently work in machine consciousness has begun to treat the phenomenon of inner speech and its possible functional role more directly (Steels, 2003). Steels' earlier work used individual-based models in multi-agent systems to investigate the development of collective lexicons. More recently he has extended these models to attempt to model syntax.

[2] Although the beeps themselves are random, statistical techniques can be used to understand the distributions of reported mental-events types and indeed correlate them with other types of behavioural measures. (R. Hurlburt & Heavey, 2004)

In Steels' newer models agents are able to check the intelligibility of their own sentences by feeding back a prospective utterance through their language interpretation machinery prior to communication. Systems of agent with such *re-entrant* loops appear to be able to self-organise more complex grammars than would otherwise be the case. (Steels, 2003, 2005) Re-entrancy in Steels' models serves the role of checking the intelligibility of an utterance in their own reception systems. Systems of such "self-talking" agents seem to be able to achieve much more stable grammars as a result.

It seems that in order to develop the abilities to use complex syntax, re-entrant loops may be necessary. Steels is thus able to persuasively link re-entrancy to the generation of complex grammars in natural language and perhaps thereby provide a functional role for the inner-voice.

One problem for this work is that the everyday construction of grammatical sentences is usually considered a largely *unconscious* activity. In fact, the construction of grammatically correct sentences is often given as the paradigmatic example of what an unconscious cognitive process is like. Thus, there seems a little *prima facie* implausibility in correlating the phenomenological inner voice with a mechanism whose principle cognitive role is the construction of grammatically correct utterances. While Steels' arguments about the role of re-entrancy in the generation of complex grammars are convincing, arguably however the link with the inner-voice is less well-made.

One important caveat should be put on this observation. Insofar as we are treating the ontogenesis of language in young children, and the problems of developing capabilities to use a language for the first time, it may very well be the case that a large portion of the child's cognitive resources taken up in assembling and comprehending sentences and possibly they are much more conscious of this. It may turn out that the kinds of activities that Steels models in his experiments might very well turn out to play a central role in the consciousness of young children, and perhaps be the trailblazers for more elaborate forms of conscious inner loops to be developed later in their lives.
A further task is to establish links between the Steels model and the account of the inner voice posited by theorists seeking to understand the re-organisation of cognition by language? Arguably his account could be made to fit with some of the recent accounts of language-for-thought that rely on the idea that language allows information to be passed between modules which wouldn't otherwise connect (cf- Carruthers, 2002). As the Steels model seems to have the language production and reception system rather separated from other forms of cognitive activity, it is difficult to say precisely how this relation could be established. Yet if the development of grammar turns out to be linked in this way to a re-entrant cognitive architecture, one can imagine how this architecture could become appropriated by other cognitive functions.

Although the Steels model offers an interesting attempt to show the functional importance of inner speech in order to stabilise the learning of grammars of certain complexity this model may be a special instance of the more general case where self-directed speech serves to scaffold and stabilise a whole range of cognitive functions. Yet could such a system also be linked to the phenomenology of inner-speech and the role of language in consciousness? More work clearly needs to be done in order to establish such a connection.

3 – A self-regulation model of inner-speech

Recent research conducted with Tony Morse (2005)[3] demonstrates how an alternative model of self-directed speech, still based on re-entrancy, might relate the inner-voice to a range of broader cognitive activities. The starting assumption for this work is that the cognitive role of language is better understood as one of sculpting or regulating cognitive activity rather than exhaustively representing the world (cf. Clark, 1996). Inner speech could here be seen as serving as a scaffold for developing and sustaining cognitive functions beyond the parsing and construction of meaningful and grammatical utterances.

In our model we compare a series of possible architectures for minimal cognitive agents which have to respond to instructions in order to fulfil externally indicated goals, i.e. moving objects around in a blocks world[4]. Our experiments compare several types of agents with differing architectures, some with word re-entrant loops and some without. All agents are implemented with simple recurrent neural networks that are evolved with a genetic algorithm in order to respond to commands by performing tasks. Some of the agents have architectures that allow the re-triggering of command reception systems internally.

The cognitive architecture of the 're-entrant' agents is arranged such that they can re-use the channels which are being used to signal commands to them to re-trigger their own behaviours. These channels allow at least the possibility of establishing new control circuits that use the same nodes that have previously been used to receive input from external 'words'. The thought here is that if there is some advantage to be had by re-using circuits developed to respond to words then the agents will take advantage of this source of useful adaptation. We find this is the case. Even such minimal agents can take advantage of these contingencies to develop word-based modes of self-regulation.

We show that agents with these 're-entrant speech' capabilities (as illustrated in **Figure 1**) perform considerably better on certain tasks. This is explained in greater detail in (Clowes & Morse, 2005). The basic finding is that agents that have architectures allowing the re-use of language for self-regulation achieve higher levels of performance more quickly and can stabilise them for longer that those that do not. Agents that are able to succeed in all task conditions make considerable use of auto-stimulation with words, i.e. they use re-entrant word nodes to self-trigger.

Re-entrance does not function in our models to facilitate merely communicative success or the generation and interpretation of complex linguistic constructions, but in the construction of more viable behaviours. Words here are appropriated in a way that is reminiscent of what Dennett calls auto-stimulation but not as a complex self-question (Dennett, 1991), but as new mode of self-regulation. This work then supplies at least a proof of concept that word-like constructs can be appropriated from a role in regulation from the outside (response to a command) to internal regulation (the agent self-regulating).

But linking such quite basic modes of auto-stimulation with words to inner speech, suggests a rather different picture of its un-

[3] A much more detailed examination of this work is now available in my unpublished DPhil thesis.
[4] NB. This is not exactly a blocks-world in the traditional sense. Rather, agents have extensive sensori-motor couplings with their limited world rather than it being specified in a purely abstract way. The agent architecture itself is an extension of an active vision model reported in experiments by (Floreano, Kato, Marocco, Sauser, & Suzuki, 2003)

derlying nature to that suggested by the Steels model. Inner speech is, I argue, the phenomenological dimension of internalised, word-based self regulation.

The phenomenological appearance of such speech, as speech, depends on it playing a similar attention focusing role as outer social speech often does. Further, I would conjecture that it relies on the same neural circuits, albeit appropriated for new self-directed functions.

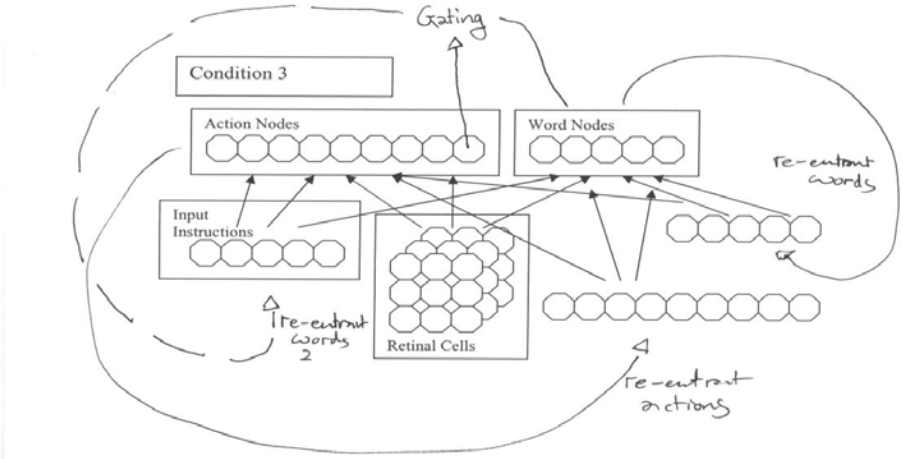

Figure 1 - The diagram shows an outline of the neural-architecture that is used in the experiments. The salient aspect is that when a gating neuron is switched on, activity from the output of the network can be fed back through the nodes that are used as input instructions. More detail on the architecture and some tasks can be found in (Clowes & Morse, 2005). Agents evolved in these conditions develop elaborate self-control loops and develop and stabilise solutions to more tasks than those that do not have such loops.

4 - A functional role for inner-speech

Normal intersubjective speech can certainly play a role in orienting attention, so why not internal speech? A shout in the street can cause an immediate refocusing of attention, e.g., hearing someone shout "mind the car!" as you were about to cross the road, would cause a fundamental reallocation of your attention.

If the inner voice could similarly be linked in some way to the allocation of attentional resources then there is the possibility that it may provide a window into the relationship between higher cognition and consciousness more generally. According to Vygotsky the internalisation of speech forms a whole new mode of attentional reorganisation.

Vygotsky (1986) emphasized the role of language in the development of control of action and ultimately of attention. His work provides an interesting possible way into the relationship between inner-speech and consciousness by looking at it through a developmental prism.

Vygotsky developed his ideas about the *internalisation* of language in part as a critique of Piaget's ideas about so-called *egocentric speech*. What Piaget called egocentric speech, and developmentalists tend to call today *private speech,* is a type of speech that children produce between the ages of about 4 and 7. It appears to be addressed toward the self and eventually seems to disappear.

For Piaget this speech occurs toward the end of his pre-operational stage and signifieds a still undeveloped ability to take, or imagine, the perspective of others. Social speech was thought to be built form this egoistic basis as children gain more experience that the point of view of others can be different (especially through argument with peers).

A longstanding controversy has arisen amongst developmentalists about the provenance and direction of this speech. Whether it is ultimately a disappearing artefact of early developmental egotism as Piaget argued in his early writings (1926), or alternatively the establishment of the bridge to linguistically controlled higher psychological function (Vygotsky, 1986 - originally 1934), either way this speech does not seem to serve a standard communicative function.

If Vygotsky's theory is correct, then inner-speech has at least its developmental precursors in this particular form of practically oriented speech found in children. If moreover inner-speech once fully internalised could come to play a role in allocating attention then this could provide a strong link between the internalisation of language and the constitution of human consciousness. Understanding inner speech may yet prove to be the royal road to understanding consciousness.

5 – Self-Restructuring through internalisation

Much of the theoretical work arguing that language plays a role in consciousness depends on the idea that language reshapes our underlying cognitive mechanisms in some way. Exactly how and to what purpose this functional re-organisation is achieved is currently part of a lively debate. The potential for re-using language as an addition to the brain's basic modes of organisation is something which is now starting to be taken very seriously in the philosophy of cognitive science (cf. Clark, 2004; Wheeler, 2004).

Dennett (1991) has argued that the development of the self-questioning form of self-directed speech is absolutely pivotal in the construction of human consciousness and its ability to sustain elaborate narrative threads. His view on this seems linked to his position that the form of human consciousness is the effect of installing what he calls a 'serial virtual machine' on parallel processing hardware. A range of accounts of the functional role of inner speech and its relationship with consciousness have also been put forward (Carruthers, 2002; Clark, 1998; Frawley, 1997) which seek to expand upon or restructure in various ways the sort of picture developed by Dennett. Although it seems possible that episodes of inner speech are epiphenomenal and fulfil no functional role in the organisation of consciousness, it is certainly too early to rule out the contrary possibility.

One can derive a further link between self-directed speech and the functional structure of consciousness from the psychopathological literature. Evidence seems compelling that the collapse of a normal inner voice in disorders such as 'thought insertion' is often correlated with catastrophic breakdowns for the organisation of individual consciousness (R. T. Hurlburt, 1993; Stephens & Graham, 2000). Disorders such as schizophrenia are sometimes theorised as control disorders and this idea gives us a way into establishing a possible link with the functional role of internalised speech (Gallagher, 2000). It points towards some quite central role for self-directed speech in the organisation of human consciousness, if not necessarily along the lines of Dennett's model.

One difficulty with this idea is that it is still very unclear at the level of sub-personal cognitive architecture how language can come to play the types of roles that are being ascribed it by the consciousness theorists. Yet there is a dearth of cognitive models that even attempt to show how such a reorganisation might happen[5]. However, it is possible to further analyse the model described above to give some insight into how attentional control through language internalisation might be established.

The model presented here gives one suggestion as to how the sorts of complex modes of self-regulation that seem bound up with human consciousness can get underway.

The simulation work with minimal cognitive agents shows that the re-use of public symbols in re-organising the ongoing activities of self can have cognitive benefits. These appear to go beyond being able to interpret and sustain more complex languages. Rather the internalisation of language in these models has more to do with the restructuring ongoing situated action.

Analysing the models further we found that the development of the ability to re-use a system of commands appears to move through broadly three control regimes.

1. Agents develop the capacity to respond to instructions. At this stage of development agents might be decribed as passive and do not use self-directed instructions very much.
2. Agents start to auto-stimulate with instruction nodes. This regime of self-control tends to produce ineffective and unstable systems of activity, (e.g. agents can sometimes perform the tasks well but very often do not).
3. Finally agents develop much more robust forms of self-control that rely on the ability to use new regimes of action made available by the self-directed loops.

Can these results be linked with Vygotsky's ideas about the establishment of new regimes of self-control through the internalisation of speech?

Vygotsky – to some extent developing the ideas of the Gestaltists[6] - argued that the development of self-directed speech was an form of self-prompting by which children come to de-centre and move themselves from one domain of situated activity (or as he might have termed it practical thought) to another. He saw this development as being centrally involved in the establishment of self-control and attention-regulation that are characteristic of human consciousness.

The work discussed above gives us a possible way of understanding the neural-dynamics underlying the establishment of this linguistic self-regulation.

6 – Inner speech and the modelling of consciousness

Notwithstanding current attempts to develop work in synthetic phenomenology (Chrisley & Holland, 1994), for now[7], hu-

[5] Despite these lacuna in more general work on cognitive modelling and the role of language some interesting work linking linguistic and cognitive function is starting to be done (Sugita & Tani, 2002). This work however encompasses quite a distinct formulation of the idea of a role for language in cognition as does the work reported here.

[6] Gestalt psychologists wrote a great deal on the problem of insight and how it was that a problem might suddenly be restructured such that it appears in an entirely new way. Kohler was one that held that tools could play a role

[7] Perhaps forever, cf, (Nagel, 1974)

man consciousness is the only type of consciousness which we know intimately. It seems unlikely that we can afford to ignore the relevance of the role of language in attempts to model it in machines, not to mention the project of building actually conscious machines.

Theorists as diverse and as historically distant as Vygotsky and Dennett have argued that self-directed speech plays a central role in the organisation and even the construction of human conscious experience. Work by Hurlburt and others appears to show that conscious experience abounds with episodes of internal speech.

If they are right and we are serious in our attempts to understand human consciousness with synthetic techniques, then we need to develop more advanced and explicit models of the role language might play in its functional organisation. The hypothesis defended here about the functional role of internalised speech is that it is a tool for the focusing or re-focusing of attentional resources.

Inner speech then appears to be of central importance because it gives an agent the capacity to restructure not just the external world but also itself. External activity in this way becomes redeployed toward inner restructuring. Simulation models such as those discussed above give us a unique mode of developing an understanding of the functional changes that underlie such a transition.

This internalisation model of self-directed speech can be used to provide an explanation of how language plays a role in creating the regimes of complex self-control and attention-regulation that are central to the sorts of consciousness that humans have (cf Donald, 2001). It does not attempt to address the question of why any experiences are conscious at all. However, it may allow us a new vantage point on their qualitative character.

According to the sensorimotor approach or 'skill theory' of conscious experience, "experience is not something we feel but something we do" (O'Regan, 2001). The character of perceptual experience, according to this theory, is given in the mastery of sensorimotor contingencies. These contingencies of self have there own governing laws just as any other complex physical system. Developing a mastery of these laws through autostimulation-with-words might be considered akin to the development of a new perceptual modality.

This mastery of the mechanisms of autostimulation-with-words affords the refocusing of one's own attention on self. This exercise of the contingencies of self can therefore be linked, more generally, to the qualitative analysis of consciousness in terms of sensorimotor contingencies (cf - O'Regan & Noë, 2001). Understanding this refocusing of attention might help us explain the uniquely human mode of the self's perceptual presence.

References

Baddeley, A. (1997). *Human Memory Theory and Practice*. Hove, UK: Psychology Press.

Baddeley, A., & Hitch, G. (1974). Working Memory. In G. A. Bower (Ed.), *Recent advances in the psychology of learning and motivation.* New York: Academic Press.

Carruthers, P. (2002). The Cognitive Function of Language. *Behavioral and Brain Sciences, 25*(6).

Chrisley, R., & Holland, A. (1994). *Connectionist synthetic epistemology: Requirements for the development of objectivity* (No. 353): COGS CSRP 353.

Clark, A. (1996). Linguistic Anchors in the Sea of Thought? *Pragmatics And Cognition, 4*(1), 93-103.

Clark, A. (1998). Magic Words: How Language Augments Human Computation. In P. Carruthers & J. Boucher (Eds.), *Language and Thought. Interdisciplary Themes* (pp. 162 - 183). Oxford: Oxford University Press.

Clark, A. (2004). Is language special? Some remarks on control, coding, and co-ordination. *Language Sciences, 26*(6), 717-726.

Clowes, R. W., & Morse, A. (2005). Scaffolding Cognition with Words. In L. Berthouze, F. Kaplan, H. Kozima, Y. Yano, J. Konczak, G. Metta, J. Nadel, G. Sandini, G. Stojanov & C. Balkenius (Eds.), *Proceedings of the 5th International Workshop on Epigenetic Robotics*. Nara, Japan: Lund University Cognitive Studies, 123. Lund: LUCS.

Dennett, D. C. (1991). *Consciousness Explained*: Penguin Books.

Dennett, D. C. (1994). The Role of Language in Intelligence. In D. C. Dennett (Ed.), *What is Intelligence*. Cambridge: Cambridge University Press.

Donald, M. (2001). *A Mind So Rare: The Evolution of Human Consciousness*. New York / London: W. W. Norton & Company.

Floreano, D., Kato, T., Marocco, D., Sauser, E., & Suzuki, M. (2003). *Active Vision & Feature Selection: Co-development of active vision control and receptive field formation. Complex visual performance with simple neural structures.* Retrieved 30 June 2004

Fodor, J. (1975). *The Language of Thought*. New York: MIT Press.

Frawley, W. (1997). *Vygotsky and Cognitive Science: Language and the Unification of the Social and Computationional Mind.* Cambridge: Harvard University.

Gallagher, S. (2000). Philosophical conceptions of the self: implications for cognitive science. *Trends in Cognitive Sciences*.

Hurlburt, R., & Heavey, C. L. (2004). To Beep or Not To Beep: Obtaining Accurate Reports About Awareness. *Journal of Consciousness Studies, 11*(7), 113-128.

Hurlburt, R. T. (1990). *Sampling Normal and Schizophrenic Inner Experience*. New York: Plenem Press.

Hurlburt, R. T. (1993). *Sampling inner experience with disturbed affect.*: Plenum Press.

James, W. (1890). *The Principles of Psychology*.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*, 435-450.

O'Regan, J. K. (2001). Experience in not something we feel but something we do: a principled way of explaining sensory phenomenology, with Change Blindness and other empirical consequences.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*.

Piaget, J. (1926). *The Language and Thought of the Child*: Routledge and Kegan Paul.

Ryle, G. (1949). *The Concept of Mind.* Chicago: The University of Chicago Press.

Steels, L. (2003). Language Re-Entrance and the 'Inner Voice". In O. Holland (Ed.), *Machine Consciousness*. Exeter: Imprint.

Steels, L. (2005). Constructivist Development of Grounded Construction Grammars.

Stephens, G. L., & Graham, G. (2000). *When Self-Consciousness Breaks*: MIT Press.

Sugita, Y., & Tani, J. (2002). A connectionist model which unifies the behavioral and the linguistic processes. In M. I. Stamenov & V. Gallese (Eds.), *Mirror Neurons and the Evolution of the Brain* (Vol. 42).

Vygotsky, L. S. (1986). *Thought and Language* (Seventh Printing ed.): MIT Press.

Wheeler, M. (2004). Is language the ultimate artefact? *Language Sciences, 26*(6), 688-710.

Playing to be Mindful
(Remedies for Chronic Boxology)

Ezequiel Di Paolo
Centre for Computational Neuroscience and Robotics
University of Sussex,
Falmer, Brighton, Sussex BN1 9QH, UK
ezequiel@sussex.ac.uk

Abstract

There is a widespread misconception among critics of the dynamical systems approach to cognition: the emphasis on embodiment and situatedness has given the wrong impression that the only cognitive activities that can be explained under this paradigm are those concerned with ongoing coping with the current situation. To say that the body is actively situated in a world is only to highlight the most fundamental aspect of all cognitive activity. There is no doubt that the dynamical systems approach has already proven immensely more successful in such cases than traditional computational approaches. Even so, as soon as we move to other, more human, cognitive performances, such as planning or imagining, we must, critics predict, return to the tenets of cognitivism/ computationalism in some updated form, or worse still, to some kind of hybrid stance. Here I briefly examine the foundations of this claim (and find there aren't really any).

On the positive side, I raise the issue of what is the best route for connecting sensorimotor and situated intelligence with (some) human styles of cognitive activity (misleadingly characterized as "decoupled"). A dynamical systems approach is already useful because it forces us to formulate the questions that traditional representational approaches felt unnecessary to ask since they answered them almost axiomatically. What is to represent? How is it possible to alter the meaning of a situation? What sort of system is a cognizer such that the world is meaningful for her? How can a cognizer act autonomously in accordance with meanings not yet established by the situation but by her own actions?

I will very briefly discuss the life/cognition continuity thesis and show how it reveals fundamental issues about agency and sense-making that allow us to begin to answer some of these questions. A powerful methodological guidance is found in Hans Jonas's work on value-generating activities and the evolutionary/historical thread of increased mediacy in cognition. Following a developmental version of this thread, a large part of this presentation will be devoted to examining pretend play (in authors such as Lev Vygotsky, Maxine Sheets-Johnstone, and Margaret Donaldson) as a particularly relevant activity for understanding how transitions to freer forms of meaning manipulation are inherently embodied and dynamical in nature. This will suggest new vistas and new challenges to synthetical approaches like evolutionary robotics.

The XML Approach to Synthetic Phenomenology

David Gamez*
*Department of Computer Science
University of Essex
Colchester
C04 3SQ
UK
daogam@essex.ac.uk

Abstract

One of the major challenges in synthetic phenomenology is to find a way of systematically describing artificial non-conceptual phenomenal states. This paper puts forward a solution to this problem that uses three different XML files to describe a machine's structure, internal states and phenomenology. The advantages of XML are that it can be read by both machines and humans, it is good at capturing hierarchical relationships between data and it can be automatically generated, analysed and archived. XML could also be a useful tool for other methods of representing non-conceptual mental content, such as content realization and ability instantiation. Furthermore, as scanning technologies develop, the XML approach could be applied to the neurophenomenology of humans, which would serve as a foundation for a more scientific psychology of both humans and machines and facilitate precise comparisons between the two. The XML approach outlined in this paper will be used to describe the synthetic phenomenology of Holland's and Troscianko's CRONOS robot that is currently under development at the University of Essex and the University of Bristol.

1 Introduction

Synthetic phenomenology is a recently emerging discipline that aims to describe the phenomenal states of artificial systems. This is essential for the monitoring and debugging of machine consciousness and it could also address concerns about the possibility of suffering in machines. This paper puts forward an approach to synthetic phenomenology that uses three different XML files to describe a machine's structure, internal states and phenomenology. The advantages of XML are that it can be read by both machines and humans, it is good at capturing hierarchical relationships between data, it can be automatically generated, analysed and archived, and it avoids many of the pitfalls and presuppositions of natural language. As scanning technologies develop it may also be possible to use XML in neurophenomenology, which would allow detailed comparisons between human and artificial systems.

The first part of this paper covers some of the limitations of natural language descriptions of the phenomenology of non-human systems. After setting out the advantages of an XML approach, the central section outlines one way in which XML representations could be used in synthetic phenomenology. This is not intended to be a final and definitive methodology, since there are no doubt better ways of applying XML to this area. However, by presenting one way in which it could be done I hope to make the case that XML could be a very useful tool for the phenomenology of artificial systems.

2. Problems with Describing the Phenomenology of Non-Human Systems

Phenomenology, especially in the work of Husserl and Heidegger, derives its significance from the claim that the phenomena we experience are as important and substantial as the physical world described by science, which is often portrayed as a secondary interpretation of the phenomena. In this way phenomenology sets itself up with an 'objective' field of phenomena that are assumed to be the same for everyone and can be unproblematically described in natural human language The problem with this approach is that these assumptions about common experience start to break down once phenomenology is applied to the experiences of infants,

animals and robots. To illustrate this problem, I will consider a short extract from Wordsworth (2004), which contains a fairly straightforward description of daffodils in natural human language:

> When all at once I saw a crowd,
> A host, of golden daffodils,
> Beside the lake, beneath the trees,
> Fluttering and dancing in the breeze.

Most people have had the experience of daffodils fluttering and dancing in the breeze and when Wordsworth's description is read by humans, they can readily imagine a similar past experience and understand his words well enough. However, even this straightforward description presents problems since it is extremely vague and imprecise and each reader will imagine the daffodils differently. More serious problems start to arise when we try to use ordinary language to describe the experiences of an infant placed in front of a field of daffodils. As Chrisley (1995) points out, we cannot simply say that the infant sees a host of golden daffodils because the infant has a preobjective mode of thought, which is unable to locate the daffodils within a single unified framework. Adults understand daffodils as something objectively located in three dimensional space, whereas infants do not necessarily continue to believe in the existence of the daffodils when they are occluded. In the adult and infant the word "daffodils" refers to two different concepts and experiences. As Chrisley puts it: "The infant's concepts are not fully objective and are therefore non-conceptual. To ascribe conceptual content to the infant in this case would mischaracterize its cognitive life and would not allow prediction or explanation of the infant's behavior." (Chrisley, 1995: 145).

These problems become even more difficult when the attempt is made to describe the phenomenology of a non-human animal, such as Nagel's famous bat (Nagel, 1974). When a bat flies over a field of daffodils it receives a complex pattern of returning ultrasound pulses, which are processed into phenomenal experiences that are likely to be very different from our own. Sentences like "the bat is experiencing a host of golden daffodils" are at best an extremely misleading description of the bat's phenomenology.

The same problems are encountered when attempting to describe the phenomenal experiences of artificial systems. Whilst we may have grounds for attributing phenomenal consciousness to some robots, we have almost no basis for believing that they will have the *human* phenomenal experience of yellow when daffodils are placed in front of them, or even that they will have the same experience of yellow as each other. Robots may also be built that have unconscious daffodil recognizers, so that they are only conscious of the abstract presence or absence of daffodils. Other robots might only be capable of processing stationary daffodils, leading to highly divergent phenomenal experiences that cannot be captured in ordinary language.

Natural language evolved to describe human experiences and so it is not surprising that it is very bad at describing the phenomenology of bats and robots. Synthetic phenomenology needs a better and more systematic way of describing the phenomenal states of artificial systems and the central claim of this paper is that XML representations are more appropriate for this task. After setting out the advantages of an XML approach, section 4 will demonstrate how it can be used to describe synthetic phenomenal experiences in a systematic manner.

3. Advantages of XML for Synthetic Phenomenology

The eXtensible Markup Language (XML) is a platform-independent way of structuring and organising data so that it can be easily shared between systems. XML is stored as plain text and it has a tightly structured format that enables the relationships between data items to be easily expressed. It is also possible to validate the structure of an XML file without any prior knowledge of its form. XML is starting to be used widely and there are a number of reasons why it would be suitable for synthetic phenomenology:[1]

1. XML is much more precise and highly structured than natural language, which allows it to describe complex nested hierarchies and represent the relationships between different pieces of information. This also enables easy cross referencing between different files.
2. XML can describe low level details of the system's hardware, but it can also abstract from them so that high level comparisons can be made between machines with different architectures and between humans and machines. Whilst two systems' lower levels might be different – perhaps using neurons or silicon - the higher levels are likely to be more similar, allowing direct comparisons between different systems once everything is encoded in XML.
3. XML can be written and read by both machines and humans. When doing simple small scale analyses it is useful to be able to manually read and edit an XML description of a machine's inner state. However, it is also very easy to automatically generate and analyse the state of a machine using XML, for example by writing programs

[1] A good XML tutorial can be found at: http://www.w3schools.com/ xml/default.asp

that look for phenomenal mental content using different theories of consciousness.

4. Its human and machine readability also make XML good for debugging consciousness. Once you have a highly structured representation of a machine's inner state and a methodology for analysing this for phenomenal consciousness, you can see how the machine's phenomenal states can be improved or increased.
5. XML is easy to archive, either by converting the XML files into a database format or by storing them directly. Sequences of mental content that are stored in this way can be examined later off-line.
6. XML is a good foundation for the other techniques for representing non-conceptual mental content, such as those suggested by Chrisley (1995).
7. XML is very flexible. In addition to tags and data, XML can contain references to external files, pieces of code and equations. This enables it to include features that cannot be precisely described in human language.

Although these advantages also apply to some of the alternatives to XML, such as JSON, YAML and OGDL, the popularity of XML and the availability of good parsers in most programming languages make it the best choice for the approach to synthetic phenomenology that I am setting out in this paper.

4. The XML Approach

This section outlines one way in which XML could provide a systematic framework for describing the phenomenology of artificial systems. This approach works using three separate but interlinked XML representations:

1) *System.* A systematic description of the system and its sensors.
2) *Test Suite*. Identifies active elements within the system that are systematically correlated with outside events impinging on the sensors. This treats the machine as a complete unknown that is systematically probed by exposing it to stimuli and measuring changes in its internal state. During the generation of the test suite no attempt is made to say what the stimuli might be like for the machine, although human descriptions are included to help with later analysis.
3) *Mental Content.* If the test suite is constructed in enough detail, a good idea should be gained about the range of correlations between internal states of the machine and activation of the machine's sensors by the outside world. However, at any point in time only a small proportion of the potentially active elements will be active and this set of currently active elements are recorded in a third XML representation of the machine's mental content. This includes tags to indicate whether it is phenomenal mental content, which are filled in at a later stage by programs designed to analyse the system, test suite and mental content XML for signs of consciousness.

The XML structures that could be used to contain the data for each of these stages will now be covered in more detail.

4.1 System

The system XML file describes the structure of the system, including sensors, actuators and internal components. This is needed to clarify the range of tests that could be applied to the system and to help with the identification of potential phenomenal states. Some extracts from an XML file describing a typical system are given below:

```
<system>
  <description>Robot</description>
  <sensor id="1">
    <type>light</type>
    <shape>rectangle</shape>
    <width>400</width>
    <length>300</length>
    <coordinate_system>Cartesian
                </coordinate_system>
    <wavelength_range>0.7-0.4
                </wavelength_range>
  </sensor>
  <!-- Add more sensors here -->

  <actuator id="1">
    <type>motor</type>
    <location>wheels</location>
  </actuator>
  <!-- Add more actuators here -->

  <neuron id="1">
    <position>2,3,3</position>
    <type>pyramidal</type>
    <algorithm>Leaky integrate and
                    fire</algorithm>
  </neuron>
  <!-- Add more neurons here -->

  <connection id="1">
    <presynaptic_neuron>1
              </presynaptic_neuron>
    <postsynpatic_neuron>3
              </postsynpatic_neuron>
    <synapse_type>excitatory
                  </synapse_type>
    <weight>0.9</weight>
    <delay>22</delay>
  </connection>
  <!-- Add more connections here -->

</system>
```

Brief explanations of some of the more important tags are as follows:

<sensor> A sensor sensitive to light, touch or sound, for example.
<**actuator**> An actuator, such as a motor or hydraulic piston.
<neuron>, <connection> In this system the internal states are held in neurons, whose parameters are specified here along with the connections between them. Other systems might use Bayesian networks or first order logic to hold their internal states.

4.2 Test Suite

A test suite is a systematic way of linking the presence of events and objects in the environment to changes in the machine's inner state. To generate a test suite the system is probed using a number of different tests and correlations between the stimulus and the machine's state are recorded as a list of active elements. The behaviour of the machine is also treated as data that is correlated with its internal states. To avoid presuppositions about three dimensional space, the input to the machine is specified in terms of changes in the machine's sensors and not as the presentation of three dimensional objects. With systems based on real or simulated neurons the test suite could be created by following the traditional approach of recording from neurons or groups of neurons. Systems along the lines of Franklin's IDA (Franklin, 1998) could be tested by using a debugger to monitor which variables or memory locations change in response to environmental stimulation. This avoids problems raised by Searle (1980) about the difference between manipulating a symbol and understanding a symbol since no assumptions are made about the meaning of any of the system's internal states.

A comprehensive test suite needs to be designed with care so that it can probe all possible sensitivities of the machine and specify them as precisely as possible. This could start with simple low level features, such as points, lines, and edges and work its way up to more abstract stimuli, such as faces and houses. All of these single modality tests would have to be combined with input from other modalities, such as audition, proprioception and sensation. They would also have to be carried out whilst the machine is engaged in different activities, such as looking to the left, moving forward, and so on, to take account of sensorimotor contingencies. Whilst this sounds like an enormous quantity of work, initial tests of this type are likely to be carried out on very simple machines and as the methodology develops it will be possible to automate the creation of the test suite by writing programs that examine the system XML file and generate a comprehensive series of tests. The tests could also be automated in many cases by simulating the input to the sensors. Some sample extracts from a test suite XML file are given below:

```
<test_suite>
  <test id="1">
    <human_description>Moving
          forward towards point of
          light</human_description>
    <sensor_input>
      <sensor>1</sensor>
      <type>light</type>
      <size>5,5</size>
      <location>55,44</location>
      <wavelength>0.55</wavelength>
      <file>Test1.dat</file>
    </sensor_input>
    <!-- Add more sensor inputs -->

    <actuator_output>
      <actuator>1</actuator>
      <type>motor</type>
      <direction>clockwise
                        </direction>
      <speed>5</speed>
    </actuator_output>
    <!-- Add more actuator outputs -->

    <active_element>
      <type>neuron population</type>
      <neuron id="27">
        <firing_rate>0.88
                      </firing_rate>
      </neuron>
      <!-- Add more neurons -->

    </active_element>
    <!-- Add more active elements -->

  </test>
  <!-- Add more tests -->

</test_suite>
```

Some of the more important XML tags are as follows:

<test> A test that is applied to the machine to probe its responses to a particular stimuli. Tests that do not activate any elements do not need to be included.
<human_description> Description of the stimulus by humans, which may be useful as part of the process of describing the phenomenology of the machine.
<sensor_input> Input is defined in sensory rather than world coordinates. This is to avoid the presupposition of three dimensional space that might be made if we talked about presenting a round object at a distance of three metres, for example.
<actuator_output> Any actions carried out by the machine whilst the stimulus is being presented.
<active_element> The part of the machine's inner state that is activated by the test. In a neural system

this could be a single neuron or a population of neurons with a particular distribution of firing rates. In a more traditional computer system this could be a list of memory locations that are altered by the stimulus. Active elements are defined in relation to the test stimuli that activated them and have no meaning outside of this context.

4.3 Mental Content

Only a small proportion of the elements inside the machine that respond to stimuli are likely to be active at any point in time. The currently active elements are stored in the mental content XML file, along with the active connections between them. This mental content is capable of influencing actions and could be involved in planning. For example, if a machine has a group of simulated neurons that selectively respond to images of houses, then these neurons could initiate motor patterns that cause the sound "house" to be emitted. The house-sensitive neurons could also become activated when the machine was offline, leading to an experience analogous to imagining or dreaming about a house. Some of this mental content may be conscious and a tag has been included to record whether this is the case. The contents of this tag are filled in at a later point when the system, test suite and mental content XML files are examined according to a particular theory of consciousness (see next section). Sample extracts from a mental content XML file are given below:

```
<mental_content id="66">
  <time>4010551056</time>
  <active_element>
    <id>2</id>
    <intensity>0.7</intensity>
    <phenomenal>yes</phenomenal>
  </active_element>
   <!-- Add more active elements -->

  <active_connection id="3">
    <type>synchronisation</type>
    <from>1</from>
    <to>2</to>
  </active_connection>
  <!-- Add more active connections -->

</mental_content>
```

Some of the more important tags are as follows:

<active_element> Reference to one of the active elements defined in the test suite along with some of its current properties.
<active_connection> An active connection could be synchronisation between firing neurons, active processing by the CPU or simultaneous broadcast along a radio link. Since active connections are not necessarily topologically bound they are defined separately from the static connections in the system file.
<phenomenal> Records whether this active element is phenomenal mental content. The contents of this tag are filled in by examining the system, test suite and mental content XML files for signs of phenomenal consciousness.

4.4 Phenomenal Mental Content

The final stage in the description of the phenomenology of the machine is the identification of the parts of the mental content that are likely to be phenomenally conscious. This is done by analysing the system, test suite and mental content XML files using a theory of consciousness. It is highly likely that different theories of consciousness will make different predictions about the phenomenal mental content of the machine, which provides a good way of discriminating between them by comparing their different predictions with first person reports about phenomenal states.[2] This process of identifying the phenomenal mental content will now be illustrated using Tononi's φ, Aleksander's axioms and Metzinger's constraints.

4.4.1 Tononi's φ

According to Tononi (2004) consciousness is linked to a system's capacity to integrate information. This is precisely quantified by Tononi as the number φ, which is the amount of effective information that can be exchanged across the minimum bipartition of a complex, where a complex is the subset of elements with $\varphi > 0$ and no inclusive subset of higher φ. Whilst there is not space to go into the details here, the system, test suite and mental content XML representations outlined in this paper would make it easy to calculate the amount of φ and pinpoint the active elements with high φ that are likely to be phenomenally conscious. It would even be possible to add a φ tag to the active elements within the mental content XML file.

4.4.2 Aleksander's Axioms

Aleksander (2003) put forward five axioms as a set of mechanisms that are thought minimally necessary to underpin consciousness. These are depiction, imagination, attention, planning and emotion. Although these axioms are not necessarily sufficient for consciousness, they are a good starting point for deciding whether a machine might be capable of conscious states and the XML approach offers a good way of analysing a system for their presence. For example, the test suite XML of an agent that

[2] There may also be ways of indirectly testing the predictions made by different theories of consciousness.

was capable of depiction would contain active elements linked to external stimuli, and an agent would be experiencing imagination when its mental content XML contained active elements that were linked in the test suite to different stimuli from the ones that are currently present. For example, an active element might be linked to apple stimuli in the test suite and yet be part of the agent's mental content when only bananas are in its field of view. One way of identifying the axiom of attention would be follow Damasio (1999) and Metzinger (2003) and look for active connections between active elements linked to the agent's self model and active elements associated with external content. Emotion could be discovered by looking for active elements associated with certain body states.[3]

4.4.3 Metzinger's Constraints

Metzinger (2003) set out eleven constraints that mental content must conform to if it is to be conscious. There is not space to go into the constraints in detail here, but the three most important, which are used to define a minimal notion of consciousness, are the activation of a coherent global model of reality (constraint 3) within a virtual window of presence (constraint 2) both of which are transparent (constraint 7). A system whose mental content conformed to these constraints would have a phenomenal experience of "the presence of one unified world, homogenous and frozen into an internal Now, as it were." (Metzinger, 2003: 169).

The identification of which parts of the mental content conform to Metzinger's constraints is easier than it seems because Metzinger provides very detailed descriptions of the informational, representational, computational and functional characteristics of the constraints along with some likely neural correlates. All of this can be fairly easily extracted once detailed and systematic XML representations have been created for the system. For example, the presence of constraint 3 (integration within a global model of reality) could be established by looking at the active connections between active elements or possibly using Tononi's methodology. Some of the other constraints, such as transparency, may come for free on systems whose internal states do not have any sensors that could make them objects of representations.

[3] The identification of planning in an agent's XML descriptions would require a fully temporalised version of the XML approach, which is not covered here.

4.5 A Description of the Synthetic Phenomenology?

Given the history of phenomenology, we might expect the final outcome of synthetic phenomenology to be a natural language description. Even if we cannot achieve this at present, it might be thought that this should be the final goal of the procedures outlined in this paper. Viewed from this perspective, the system, test suite and mental content XML would only be the preparatory stages for a traditional phenomenological account of the experiences of COG, CRONOS or IDA.

However, the problems discussed in section 2 make it unlikely that we are ever going achieve fluid natural language descriptions of non-human systems. Instead, it might be much better to treat the XML representations as the best description that we are going to get of the phenomenology of an artificial system. This has the great advantage that it is possible to see what you cannot say. We don't have adequate words in human language to describe a system that can only experience vertical lines, but we can represent such a system accurately using XML, and by looking at the XML we can start to understand how much and how little we can imagine what it is like to be such a system.

The XML descriptions also offer a good starting point for other ways of describing the phenomenology of artificial systems. The suggestions made by Chrisley (1995) about conceptual subtraction, content realization, ability instantiation and self instantiation could all be implemented automatically once the XML formats have been defined. XML would also enable precise comparisons with humans that have deficiencies in the same areas as a machine, and we could use the first person descriptions of these patients to help us imagine what it is like to be such a system. As scanning technology improves, the application of this approach to normal and brain damaged patients will become easier. Research by Kamitani and Tong (2005) on neurophenomenology using combinations of voxels suggests that it might even be possible to start this work today.

5. Discussion

One of the first issues that must be clarified about the XML approach to synthetic phenomenology is that it makes no presuppositions about whether any particular machine is the sort of system that is capable of supporting conscious states. Robots, stones and human beings are all systems that are capable of internal states; all three can be analysed using the XML approach that I have set out here and it will be an empirical outcome of this approach if it turns out that the mental content of a stone is always devoid

of phenomenal states. This *empirical* outcome must be distinguished from the *a priori* question about whether certain types of non-human system are capable of supporting conscious states, since it is possible that the XML approach will make predictions about consciousness in systems that we consider highly unlikely to be capable of consciousness – the economy of Bolivia, for example. This *a priori* question is tackled by the ordinal probability scale, set out in Gamez (2005), which evaluates the likelihood that a machine can support phenomenal states by systematically comparing its architecture with the human brain.

It has been suggested that this XML approach to synthetic phenomenology ignores behavioural criteria of consciousness, such as reports that a system might make about its mental contents. If this was thought to be important, then it would be easy to include the actuator outputs in the mental content XML file, so that the external behaviour of the system could be included in the analysis of its consciousness on a moment to moment basis. However, the problem with behavioural criteria for consciousness is that apparently conscious behaviour can be generated by systems that we are reluctant to attribute consciousness to (such as the population of China communicating with radios and satellites), which is why an internal architecture approach has been favoured here.

As this methodology develops there are likely to be a large number of ambiguities about what constitutes an element, how to handle overlapping elements, how to define active connections, the best way to analyse mental content for phenomenal states, and so on. Although these might initially appear to be weaknesses of the method, they are actually strengths because they indicate that synthetic phenomenology has the potential to become a paradigmatic science that can move forward by asking questions and resolving ambiguities such as these. At the moment synthetic phenomenology is so unclear that even its lack of clarity is unclear to it and tightening up the methodology through XML representations would make it capable of asking and answering precise questions and enable it to move forward in a sustainable manner. Different ways of resolving the ambiguities will make testable predictions about the phenomenal states of a machine or organism and as neural scanning becomes better we will actually be able to test these predictions on human beings and eliminate inaccurate methods. In the early stages it is likely that different theories will generate conflicting XML representations. However, this will at least make differences explicit; whereas at present our descriptions of inner states are so woolly and imprecise that disagreement or comparison between methods is rarely an issue.

For reasons of brevity and clarity this paper has set aside questions about the temporal nature of phenomenal experience. One solution to this would be to break the stimuli up into sequences of frames and separate the test suite and mental content into a list of associated XML files. Another temporal problem is that active elements may change as they develop and so it may not be possible to generate a single test suite that is valid for all time. This type of system will have to be retested at regular intervals or have its adaptivity frozen whilst the description of its synthetic phenomenology is taking place.

6. Previous Work

The approach that I have set out in this paper is closest to some of the techniques for representing non-conceptual content discussed by Chrisley (1995). These include content realization, in which content is referred to by listing "perceptual, computational, and/or robotic states and/or abilities that realize the possession of that content" (Chrisley, 1995: 156), ability instantiation, which involves the creation or demonstration of a system that instantiates the abilities involved in entertaining the concept, and two forms of self instantiation, in which the content is referred to by pointing to states of oneself or the environment that are linked to the presence of the content in oneself. Whilst all of these techniques are promising ways of referring to non-conceptual content, it will be very difficult to apply them in practice without a precise way of representing and organizing the computational, and/or robotic states and/or abilities. It is here that XML would be a useful tool since it could represent the structure of the systems that are being analysed along with their inner states when they are exposed to stimuli from the environment. Within the precise framework offered by XML the specification of non-conceptual mental content using Chrisley's techniques would be made considerably easier.

Other related work includes the description of the synthetic phenomenology of Khepera robots by Holland and Goodman (2003) and Stenning, et. al. (2005). In these experiments the internal model of the Khepera is held in a neural network, which stores a linked series of concepts combining sensory and motor information. The synthetic phenomenology of the Khepera is carried out by plotting a graphical representation of the sequence of sensations and movements stored in the neural network. The problem with this approach is that the Khepera is likely to have no notion of colour and a very limited idea about space and so this graphical representation is unlikely to be anything like the Khepera's actual 'mental' content. Another problem is that the graphical representation contains the complete in-

ternal model, whereas only a small part of this would be active at any point in time. It is also hard to see how this representation of an internal model could be systematically analysed for signs of consciousness. The XML approach could help with these problems since it offers a highly structured way of representing the current mental content of the Khepera, which could be compared with other robots and systematically analysed for signs of consciousness .

7. Conclusion

This paper has briefly outlined an XML approach to synthetic phenomenology in which XML plays a key role in the description of the conscious and unconscious states of the machine. This has many advantages and could help to circumvent many of the problems associated with the representation of non-conceptual mental content. By describing mental content this concretely it also forces us to face challenging theoretical and methodological questions, which will eventually open up the possibility of a systematic science of synthetic phenomenology that can pose and answer precise questions about the phenomenology of artificial systems.

The XML extracts included in this paper are intended as simple examples to illustrate the main ideas and a great deal more work is needed to turn these starting points into a usable method. Some of this development will be done as part of the work on the CRONOS robot at Essex and Bristol. In the longer term it may be possible to develop a single XML standard for both synthetic and neuro- phenomenology, which would facilitate precise comparisons between humans, animals and machines and enable us to automatically examine all three for signs of consciousness.

Acknowledgments

Many thanks to Owen Holland for feedback about this paper. Thank you also to the EPSRC for funding this work (grant number GR/S47946/01).

References

Igor Aleksander and Barry Dunmall. Axioms and Tests for the Presence of Minimal Consciousness in Agents. In Owen Holland (ed.), *Machine Consciousness*, Exeter: Imprint Academic, 2003.

R. J. Chrisley. Taking Embodiment Seriously: Non-conceptual Content and Robotics. In Kenneth M. Ford, Clark Glymour, & Patrick J. Hayes (eds), *Android Epistemology*, Menlo Park/ Cambridge/ London: AAAI Press/ The MIT Press , 1995.

Antonio, R. Damasio. *The Feeling of What Happens*. New York, San Diego and London: Harcourt Brace & Company, 1999.

S. Franklin, A. Kelemen and L. McCauley. IDA: a cognitive agent architecture. *IEEE International Conference on Systems, Man, and Cybernetics*, 3: 2646-2651, 1998.

David Gamez. An Ordinal Probability Scale for Synthetic Phenomenology. In R. Chrisley, R. Clowes and S. Torrance (eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment* 85-94, 2005.

Owen Holland and Rod Goodman. Robots With Internal Models. In Owen Holland (ed.), *Machine Consciousness*, Exeter: Imprint Academic, 2003.

Y. Kamitani and F. Tong. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience* 8:(5) 679-685, 2005.

Thomas Metzinger. *Being No One*. Cambridge Massachusetts: The MIT Press, 2003.

Thomas Nagel. What is it like to be a bat? *The Philosophical Review* 83: 435-456, 1974.

J. Searle. Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3: 417-57, 1980.

J. Stening, H. Jacobsson and T. Ziemke. Imagination and Abstraction of Sensorimotor Flow: Towards a Robot Model. In R. Chrisley, R. Clowes and S. Torrance (eds.): *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment* 50-58, 2005.

G. Tononi. An Information Integration Theory of Consciousness. *BMC Neuroscience* 5:42, 2004.

William Wordsworth. I Wandered Lonely as a Cloud. In Stephen Gill (ed.), *Selected Poems*, London: Penguin, 2004.

The Embodied Machine: Autonomy, Imagination and Artificial Agents

Nivedita Gangopadhyay[*†]
[*†]Institut Jean Nicod
1bis, avenue de Lowendal, 75007, Paris, France
Nivedita.Gangopadhyay@ehess.fr

Abstract

The embodied and enactive approach to consciousness emphasises the role of the physical embodiment of naturally intelligent agents as crucial for a study of consciousness and the importance attributed to the body also tends to be carried over to the material out of which the body is created viz. "living" matter. This seems to put into doubt the relevance of the embodied and enactive approach to the field of machine consciousness. However, I shall argue that consciousness as manifested in embodied intelligent systems, natural or artificial, that enact their world of experience by interacting with the environment necessarily needs to be understood in the light of freedom/autonomy and imagination, and the application of the principles of embodiment and enaction in the light of these notions in the field of robotics and AI can be a big step towards creating conscious artificial agents.

1 Introduction

The attempt to understand cognition and consciousness by recognising the fact that they necessarily involve an embodied agent who enacts her world of experience by real-time interaction with a real-world situation has been propounded of late in an ever-increasing volume of literature in the field of consciousness studies. The emphasis laid on the notions of embodiment of the cognitive agent and her interaction with the environment as crucial elements even for a scientific study of consciousness, has come a long way from a philosophical idea first presented in continental philosophy in the works of philosophers like Husserl (Husserl, 1931, 1960) and Merleau-Ponty (Merleau-Ponty, 1962, 1963, 1964). When the ideas of these philosophers were being introduced to the philosophical analyses of consciousness, mainstream cognitive science in general had remained unaffected by the implications of such a phenomenological approach. The applications of the emerging principles of cognitive science in the field of robotics and artificial intelligence dominated by the information-processing view of cognition had largely ignored the possible implementations and crucial insights that a primary emphasis on the notions of embodiment and enaction could have provided. The necessity to stress the agent's particular psycho-physical apparatus and the real-time interactions of the agent with the real-world environment for an adequate study of consciousness began to be realised for the first time in robotics and AI in the 1980s in the work of Brooks (Brooks, 1986, 1991, 1993, 1994). The development of what has come to be known as the autonomous-agent theory in AI emphasises that as a first step for artificial agents to exhibit mental characteristics typically associated with conscious agents, they must be created in such a way that they are capable of moving about, surviving and performing specific goal-directed actions in real time in a complex real-world environment.

The embodied and enactive theories, as are gradually gaining ground in mainstream cognitive science, emphasise the kind of body the agent possesses as one of the first crucial elements to be considered by a satisfactory theory of consciousness and the importance attributed to the kind of body also tends to be carried over to the material out of which the body is created viz. biological matter. The environment of the agent, both natural and socio-cultural, also constitutes an indispensable dimension of embodied and enactive approaches to the study of consciousness. Indeed, a survey of the literature reveals that "embodiment" can be understood in a variety of ways and following Ziemke we can enumerate the different notions of embodiment as follows: 1) structural coupling between agent and environment, 2) historical embodiment resulting from a history of agent-environment interaction, 3) physical embodiment, 4) 'organismoid' embodiment i.e. organism like body and 5) organismic embodiment

of autopoietic living systems (Ziemke, 2001). However, for the present I shall consider the notion of embodiment in a rudimentary sense of physical embodiment i.e. having a particular kind of body as a fundamental determinant of consciousness. Having a kind of body means instantiating a specific biological model and a major contention of the embodied and enactive approaches is that the biological model of the agent crucially determines the characteristics associated with consciousness exhibited by the agent. While this claim is indeed justifiable in view of the fact that the interaction of the organism with the environment that generates its world of experience is importantly determined by the physical embodiment of the organism, the implication that all manifestations of consciousness could thereby also be limited to the embodiments of naturally intelligent systems as created out of "living-matter" is less evident. Due to this insistence on the physical embodiment of the agent and the underlying importance of the biological matter, theories of embodied and enactive cognition seem to possess an unavoidable "biological flavour". Then does engineering artificial intelligence according to the basic principles of embodiment mean engineering living matter? This may seem to stand in the way of applying the principles of the embodied and enactive approach in the field of machine consciousness. In this article I would like to address the question: *How far can the embodied and enactive approach to consciousness with its emphasis on the kind of body possessed by the naturally intelligent agents at all help in understanding consciousness through the creation of intelligent machines based on the principles of embodiment and enaction?* I shall maintain that although the notions of embodiment and enaction as used for understanding naturally intelligent systems importantly involve the material out of which the agent's body is created as a determinant of embodiment, these notions can be applied in the field of robotics and AI too to create artificial agents that we could at least hesitate to call "machines" even if the material out of which they are created is not "living" stuff. I shall argue that consciousness as manifested in embodied intelligent systems, natural or artificial, that enact their world of experience by interacting with the environment necessarily needs to be understood in terms of freedom/autonomy and imagination, and the application of the principles of embodiment and enaction in the light of these notions in the field of robotics and AI can be a big step towards creating conscious artificial agents.

1.1 Natural Embodiment

The embodied and enactive theories have at times sought to differentiate between naturally intelligent systems and mechanical systems by drawing upon the material out of which each is created and the set of structural properties of the resulting systems as the criteria. One such effort is made by Maturana and Varela (Maturana and Varela, 1980, 1987) who distinguish between autopoietic systems and allopoietic systems. Biological systems, made out of living matter and exhibiting natural intelligence, are essentially characterised by their adaptability to their environment at the cellular as well as at the behavioural levels. Such systems are termed autopoietic as they are self-creating and self-maintaining systems, and hence are completely autonomous. On the other hand, mechanical systems made out of non-living matter are capable of adapting only at the behavioural level and are called allopoietic systems i.e. systems whose components are produced by other processes that are independent of the organization of the machines. Hence how can artificial agents created out of non-living matter help us understand consciousness? Here one may adopt a stance of mysterianism and claim following Prinz that "…progress in the science of consciousness may offer little help to those who want to engineer consciousness" (Prinz, 2003) because it is impossible to determine with certainty that biological matter does not contain properties essential for consciousness. Then are our efforts to employ the principles of embodied cognition to the study of robotics futile unless we make machines out of living matter? Prinz advises engineers that they should not "…fool themselves into thinking that they can definitely create conscious machines" (Prinz, 2003) and the emphasis laid by the embodied cognition approach upon the body and of what it is made may seem to lend support to Prinz's advice to engineers. Thus of what use are the notions of embodiment and enaction in the study of robotics and AI?

Given our present state of knowledge about biological matter we cannot but maintain for the time being that it is in fact impossible for us to determine with complete certainty that organic matter does not contain properties essential for consciousness. However, this does not make the notions of embodiment and enaction a redundancy for robotics. Instead of considering the problem of consciousness in its totality, in all its aspect, let us begin by picking out a feature that can be said to be invariably associated with manifestations of consciousness in agents with a biological embodiment. Naturally embodied agents constantly strive to attain to higher degrees of freedom by actively resisting and defying the various forces acting against them that try to break up the unity of the system and by such efforts they assert their existence. The more they are able to resist the counteracting forces threatening to destroy the unity of the system, the more they appear to be complex

from the point of view of consciousness. Thus the most rudimentary life-form embodied in the simplest biological embodiments is the least able to actively preserve its unity in the face of counteracting forces and possesses the least freedom from this point of view and is ascribed the least traces of consciousness. As we go higher up the evolutionary chain we find more and more complex life-forms with more and more complex embodiments with greater and greater degrees of freedom exhibited by actively resisting counteracting forces till we reach the human level to which we ascribe the highest intelligence and consciousness exhibited so far in the story of evolution. Moreover, over and above adverse natural forces biological systems also deal with highly complex socio-cultural forces even at a low level of the evolutionary ladder, e.g. complex social structure of termite or ant colonies, and they seek to maintain their individual existences in this social maze by variously manipulating the forces at work there and trying to preserve their identity as individuals i.e. the unity of their individual systems. When they try to preserve and assert the identity of a group they do so as they identify the unity of their systems with that of the group. The more complex the forces that act against the system and the more the system tries to exert its freedom in the form of preserving its unity by actively counteracting the forces, the more it seems to manifest intelligence. From an evolutionary perspective it can be said that the forces acting against the system become more and more complex as we go higher up the ladder and the ways of counteracting those forces also become more and more sophisticated and complex leading to the expressions of greater freedom and accordingly greater degrees of consciousness.

As a primary strategy of counteracting the forces acting against the system naturally intelligent agents resort to interacting with their environment in creative ways i.e. *they can represent to themselves or enact possible states of affairs by interacting with the present state of affairs*. The ability to represent possible states of affairs varies in complexity in accordance with the embodiment of the system and the complexity of the forces which the system encounters. The human form of embodiment is the one most capable among all biological embodiments to actively maintain the unity of its system in the face of highly complex counteracting forces, both natural and socio-cultural; and the capacity of humans to enact possible worlds, as a strategy adopted for counteracting adverse forces, is remarkable among biological embodiments from the point of view of its complexity. This capacity as present in human embodiment is what we generally call *imagination i.e. the enaction of possible worlds* although other naturally intelligent systems too can represent to themselves possible states of affairs in various degrees of complexity by interacting with the immediate environment (the present state of affairs) and hence can also be called "imaginative". By this remarkable capacity/strategy of counteracting disintegrative forces naturally intelligent systems exert their greatest freedom.

Moreover, in case of naturally intelligent systems it can be observed that with the increasing complexity of embodiment the interaction of the organism with the environment gradually shifts from one of adaptation to one of gradual control leading to greater expressions of freedom and intelligence. The simplest life-forms adapt themselves as best as they can to the conditions of the environment and accordingly the manifestation of intelligence in them is far less complex than that of the higher ones. The strategies of interacting with the environment tend to become more of control and less of adaptation in more and more complex embodiments till we reach the human level that is crucially characterised by its capacity to enact possible worlds by interacting with the environment primarily in the form of *control strategies*. In case of natural forces humans do not submit themselves to the mercy of Nature and try to adapt as best as they can to the situations Nature throws them into. Humans exert their freedom against natural forces by trying to master natural laws and make them work for their best advantage. Even for socio-cultural forces humans demonstrate the tendency to assert their control over the environment and this tendency has been manifest throughout the history of human civilization. By interacting with the environment (the current state of affairs) in accordance with their embodiments humans enact possible states of affairs (imagination) that can be far removed from and greatly more complex than the present state of affairs. Hence in humans, manifestations of intelligence are not simply matters of adaptation; intelligence is dominance and control over environment with the aim of manipulating it to the best of their advantage i.e. making conditions most favourable for the maintenance of the unity of the system and thereby exerting their freedom. For other biologically embodied systems too intelligence is crucially determined by the ability of the system to actively preserve its unity in the face of counteracting forces and by interacting with the environment in creative ways to represent to itself possible states of affairs and thereby asserting its freedom. No matter how simple or how complex the embodiment, the basic principle of intelligence and manifestation of consciousness indeed seems to be this and the human embodiment by virtue of the greatest ability exhibited so far in evolution to preserve the unity of the system and enact possible worlds by interacting with the present environment, enjoys the greatest degree of freedom in the chain

of evolution and exhibits the most complicated manifestations of intelligence.

Furthermore, despite varying in degrees of complexity and freedom, the naturally intelligent systems are all characterised by the ability to represent to themselves the goals of their actions. Natural systems have *dynamic needs* and so they interact with the environment in various ways and enact their world of experience. Exploration of the environment by natural agents is importantly guided by *curiosity*, i.e. the *need* to explore more. This is all the more true for human agents whose insatiable curiosity has been at the root of all discoveries and inventions. The interaction with the environment by human agents is characterised by this need to explore more and more, and the lack of complete satisfaction with the present state of affairs. The need to explore more is developed by the system by means of interacting with the environment and by representing to the system goals other than the immediate ones present in the environment i.e. enacting possible states of affairs (imagination). The biological systems express their freedom by not being limited to what is immediately present in the environment. The needs, whether biological or psychological, can be traced back to the desire to assert the existence of the organism and preserve the unity of the system and enact possible worlds by interacting with the present environment. Thus the needs come from the system by interacting with the environment and as long as there is embodiment there are needs.

2 Autonomy, Imagination and Artificial Agents

Consciousness as associated with this idea of freedom expressed by the embodied system through actively resisting disintegrative forces to maintain the unity of the system and enacting possible states of affairs by interacting with the present state of affairs need not be logically restricted to biological systems alone. This idea can be applied to the study of robotics and AI although the material out of which we create artificial agents like robots is not organic matter. The question is one of freedom. Naturally intelligent systems are characterised by various degrees of freedom in that they have capacities to actively preserve the unity of their system against disintegrative forces in various degrees and enact possible states of affairs by interacting with the present state of affairs, and accordingly manifest various degrees of complexity of intelligence. However, while modelling consciousness it is to be noted that biological embodiments, including human embodiment, have been shaped primarily by the environment whereas for artificial agents it is humans who are *exclusively* trying to shape the embodiment. The application of the embodied approach to cognition has to date influenced the shaping of the embodiment of artificial agents in so far as engineers are now trying to derive inspiration from biological models. The creation of robots that simulate the embodiment of simple biological models like insects (Beer and Chiel, 1993) etc. reflect this urge to copy Nature's work. This is certainly a big step towards realising the importance of the embodiment and enaction for consciousness but it is one thing to mimic biological models for embodiment of artificial agents and quite another thing to create artificial agents whose embodiment will be shaped by the environment, which includes humans but *not only* humans, by means of *creative interaction of the system with the environment* and in order to creatively interact with the environment the system must be able to *develop its own dynamic needs*. To quote Ziemke, " …despite all biological inspiration, today's adaptive robots are still radically different from living organisms. In particular despite their capacity for a certain degree of self-organization, today's so-called 'autonomous' agents are actually far from possessing the autonomy, and consequently the embodiment of living organisms." (Ziemke, 2001). Thus the idea of autonomous agents, that is already prevalent in the study of robotics under the influence of the ideas of embodiment and enaction, can be carried to a greater extent to create agents which are autonomous not only in so far as they are capable of acting upon the environment to carry out functions that have already been decided for them such as moving about and avoiding obstacles but to create systems that will develop their own course of actions by interacting with their environment in accordance with *their dynamic* needs.

To further clarify the idea let us consider basic applications of the idea of embodiment in the field of machine consciousness such as Brooks' "mobots"(Brooks, 1986, 1991, 1993, 1994). Brooks lays down four conditions that his artificial creatures should satisfy and one of these is that a creature must do something in the world; it should have a purpose in existing (Brooks, 1991). Thus Herbert, one of Brooks' well-known mobots, was designed to collect empty soft-drink cans left in the MIT lab. Although Herbert was built on the principles of interaction with the environment, it was none-the-less the human factor that exclusively fixed Herbert's reason for existence and limited its activities in important ways and thereby the exhibition of intelligent behaviour on its part. To understand this more clearly let us compare a human agent with Herbert performing the same task i.e. collecting empty soda cans in a lab. The ways of navigating through the real-world environment maybe quite similar for both the agents but the *reasons* for doing so are crucially different in case

of the human agent and Herbert. The human agent can be collecting cans by taking part in an experiment or by being employed by the lab or because she cannot tolerate a messy littered lab or simply because she likes to collect cans. However, in all these cases she knows that collecting cans is not the reason for her existence; she can stop collecting them (at least in her mind) if she wants and this representation of a possible state of affairs is an important component in her performance of the task. Even if she has been employed by the lab to collect cans she can conceive of possible worlds where she is not conditioned to collect cans. If she is bored or tired with the task or if she simply thinks it has been enough for her she can just quit. That is to say that by means of interaction with this environment i.e. the lab (the present state of affairs) she can enact a possible state of affairs. The autonomy expressed by the human agent in the task is that she is free (at least in her cognitive world) *to make a choice*; to collect cans or not to collect cans? This autonomy importantly shapes the way the human agent interacts with the environment even for simple tasks such as can-collection. Enaction of possible worlds is significantly determined by the interaction with the present state of affairs or the immediate environment. How can this autonomy be brought into artificial agents? To answer this question I shall make use of the notion of *potential enactive state* in the modelling of consciousness.

In creating an artificial agent in accordance with the principles of embodiment and enaction it is necessary to build in some routines in the form of reaction to the various environmental factors. For example, the subsumption architecture underlying the functioning of Herbert is composed of layers which can be viewed as built-in routines of reactions to environmental factors like halting when an object is sensed right in front and reorienting towards an unobstructed direction. It is crucial for successful elementary navigation through the environment that certain rules of interaction with the environment be present in the robot that guides its behaviour. These can be considered as routines that enable the artificially embodied agent to enact the present state of affairs by interacting with the environment. However, the cues that the robot obtains by interacting with the environment need not all be directed towards solving a specific task either in the form of positive feed-back or negative feed-back. Imagine a device that has multi-sensors simulating the senses of natural agents. The inputs that the robot receives via its interaction with the environment need not all be translated into action. Some inputs will be utilised for immediate action whereas some will not be so utilised. However, the ones that are not so utilised immediately will not be ignored as irrelevant for all times but be preserved in the system as potentially relevant cues for further interacting with the environment. Although the robot can be initially programmed for performing a specific task such as navigating through a real-world environment and avoiding obstacles, the picking up of cues from the environment by interaction should enable the system to develop further goals i.e. further *needs* for interacting with the environment. Interaction with the environment is a crucial factor in the origination of goals for embodied systems and the setting forth of these goals and representing them to the system enables the system to evolve and exhibit more complex intelligent behaviour. A human agent navigating through an environment for initially performing a specific task, e.g. soda-can collecting, picks up a lot of cues from the environment that are not all immediately pertinent to the task at hand but which significantly determine the manifestation of intelligent behaviour on the part of the agent. Suppose while collecting the cans in a lab a human agent hears a strain of music coming from somewhere. The music is rather lilting and the agent feels the need to dance to its tune i.e. move her body to its rhythm. The music does not constitute any part of the pertinent cues for soda-can collecting but it does constitute a dimension of interaction of the agent with the environment and enables the agent to enact a possible state of affairs. If the agent is not restricted by the terms and conditions of employment or experimentation, she may even abandon her task of can collection for some time and just dance a bit or listen more intently to the music, and if she wishes she can give up the activity of collecting soda-cans in favour of enjoying herself. Moreover, she is most likely to become curious about the source of the music too and may leave her immediate environment to trace it, i.e. she will explore more. If her movements are restricted by terms of employment or experimentation, she can none-the-less enact a possible state of affairs in her cognitive world where she is executing her desired behaviour. Thus the human agent exerts her autonomy by preserving the unity of her system in the face of counteracting forces (obligation to collect cans despite the reluctance to do so) and enacting a possible state of affairs (dancing, listening with greater attention to the music, exploring the environment for the source of the music) by interacting with the present state of affairs (collecting cans but there is a nice music coming from somewhere). In fact it is this feature of naturally intelligent systems, especially human ones, that has so far distinguished them from machines or mechanical behaviour as has been so far modelled. Herbert can go on collecting soda-cans indefinitely for it does not develop any further needs by interacting with the environment but a naturally intelligent system will sooner or later call it a day. As Maturana points out, "…as living systems that live humanly we are different from

robots on two fundamental accounts: one, is that robots have been designed de novo, intentionally in congruence with a specified medium that may also have been designed with them, and are not the arising present of an evolutionary history; two, is that we human beings are the arising present of an evolutionary history in which our ancestors and the medium in which they lived have changed together congruently…" (Maturana, 2005)

The idea of potential enactive state is the idea of the system's ability to represent to itself goals other than the ones immediately present in the environment by interacting with the environment. According to its embodiment the artificial system should be able to develop its needs with the aim of exerting greater control over the environment or for moving from adapting to the environment to gradually controlling it to the best of its advantage. The human control in the creation of truly autonomous artificial agents is at least for the time-being importantly present at the levels of design and programming. At the level of designing the initial embodiment the human designer needs to make a choice of environment for the artefact and equip the system with means of interacting with the environment. For this inspiration can be derived from biological systems and their sensory modalities because these systems by interacting with the environment for a long time have developed the most practical designs. The choice of the number of sensory modalities with which the artificial agent is to be equipped and the kind of movement that the system will execute are the concerns of the human designer, although the movement may be importantly determined by the physical features of the environment chosen. The complexities of the sensory modalities and the movement will significantly determine the complexity of the initial embodiment. However, a truly autonomous system, albeit constituted of non-living matter, should also be able *to evolve* its embodiment in accordance with the interaction with the environment. This is not merely a question of adapting to the environment at the level of behaviour as Maturana and Varela state for allopoietic systems (Maturana and Varela, 1980, 1987). It is developing or *evolving* the embodiment in accordance with environmental interaction with the aim of expressing greater freedom of the system and exerting greater control over the environment. As an example we can consider Herbert once again in an imaginative thought experiment that could roughly capture the implementation of the idea of an essential manifestation of consciousness as the ability of the system to actively preserve its unity in the face of counteracting forces and exert its freedom by interacting with the environment in creative ways to represent to itself or enact possible states of affairs. Herbert is designed to collect empty soda-cans in a lab and explores the environment randomly, *not ignoring* other objects but exploring them too by means of tactile and visual modalities. This can indeed be possible as Brooks claims that the artificial creature should be able to maintain multiple goals (Brooks, 1991). However, with the kind of physical embodiment (design) Herbert has been initially given it can pick up only empty soda-cans. Nevertheless Herbert can send a "distress" signal when it "senses" that the system is missing something, i.e. the system has developed a need by interacting with the environment and this need needs to be fulfilled for the system to exert greater freedom and control over the environment. Also suppose Herbert is equipped with temperature sensors that enable it to estimate how much energy is being spent. Now Herbert is moving through the lab, exploring it and picking up empty cans when it finds one. In the course of its random exploration suppose Herbert comes across a piece of crumpled paper lying on the floor. By exploring that crumpled ball of paper Herbert finds out that the weight of that object is less than the objects that it has been picking up. Hence interacting with that object, rather than with the empty soda-cans, means less spending of energy by the system which means more ability to explore the environment. But with the current design Herbert cannot pick up the ball of paper. It sends out a "distress" signal to indicate that the system needs something. This indicates that the agent is developing its own needs and enacting to itself a possible state of affairs by interacting with the present state of affairs for preserving the unity of the system. However, enacting a possible state of affairs or the representation of a potential enactive state should not come to an end with only a single instance. By encountering the ball of paper the system should be able to represent to itself the general possibility that there are objects in this environment that put less demand on its energy and consequently interacting with them means more ability to interact with the environment and preserve the unity of the system. *The representation of this possibility should never be exhausted.* For naturally intelligent systems as long as there is embodiment there are needs for which the system manifests intelligence and artificial embodied systems must also follow in their steps. The potential enactive state in an artificial agent represents a state which is *actually never reached* by the system. It is a state which the system is always *trying* to reach and with this aim is interacting with the environment. The system *must never reach equilibrium* i.e. the state where the system "feels" no more need to interact with the environment or develops no further need to interact with the environment. Real-world environments are essentially dynamic set-ups and hence complete control of the counteracting environmental forces is a dream for both naturally intelligent systems and

artificial ones. Yet it is the incessant pursuing of this practically unattainable state of autonomy that leads intelligent systems to manifest more and more complex intelligence.

To sum up, the necessity of applying the principles of embodiment and enaction in the field of robotics and AI is becoming increasingly clear for creating artificial agents that can exhibit mental characteristics typically associated with consciousness, and the notions of autonomy (exerting greater and greater degrees of freedom by the ability to preserve the unity of the system in the face of counteracting forces) and imagination (enacting possible states of affairs by interacting with the present state of affairs or the immediate environment) are crucial for creating embodied artificial agents capable of enacting cognitive states. The goal to be attained is complete autonomy obtained by constant enactment of possible states of affairs through interaction with the present state of affairs, and the ever present vision of this impossible goal necessarily permeates all intelligence and evolution, from the simplest to the most complex till date. I have argued in this paper that such a manifestation of consciousness need not be restricted to naturally intelligent systems alone and can be simulated in artificial agents. Whether or not it is impossible to satisfactorily determine the issue of biological matter possessing properties essential for consciousness, or whether or not autopoietic systems are essentially different from allopoietic ones by virtue of their adaptability, are questions that tend to restrict the notions of embodiment and enaction to a level of explanation that may render these notions inapplicable *in principle* in the domain of robotics and AI because of their explicit or implicit harping on biological matter. This is not to imply that these issues can be brushed aside in studies of embodiment and enaction. It may indeed be possible that biological matter is really some thing quite special for manifestations of consciousness, and the latter is inseparably linked to the former *and only* to the former. But nevertheless it may also be possible, by understanding embodiment and enaction as pertaining to consciousness in the light of freedom and imagination, to create artificial agents that we would hesitate to call "machines" any more in the sense that they perform only dumb repetitive behaviour in order to serve *our* purposes and *our* whims. Thus the challenge that faces us for this new vision of artificial agents is not how far *could* we go in creating conscious machines but rather: How far would we *dare* to go?

References

Anderson, M.L. Embodied cognition: a field guide. *Artificial Intelligence*. 149: 91-130, 2003.

Beer, R., and Chiel, H. Simulations of cockroach locomotion and escape. *Biological Neural Networks in Invertebrate Neuroethology and Robotics*. ed. R. Beer et al. Academic Press, 1993.

Brooks, R. A robust layered control system for a mobile robot *IEEE Journal of Robotics and Automation RA-2,* 1, April: 14-23, 1986.

Brooks, R. Intelligence without reason. *Proceedings of the 12th International Joint Conference on Artificial Intelligence*. Morgan Kauffman, 1991.

Brooks, R. A robot that walks: Emergent behaviors from a carefully evolved network. *Biological Neural Networks in Invertebrate Neuroethology and Robotics*. ed. R. Beer et al. Academic Press, 1993.

Brooks, R. Coherent behavior from many adaptive processes. *From Animals to Animats 3*. ed. D. Cliff et al. MIT Press, 1994.

Brooks, R., and Maes, P. eds. *Artificial Life 4*. MIT Press, 1994.

Brooks, R., and Stein, L. Building Brains for Bodies. Memo 1439, Artificial Intelligence Lab, Massachusetts Institute of Technology, 1993.

Chrisley, R., and Ziemke. T. Embodiment. *Encyclopedia of Cognitive Science*. Macmillan, 2002.

Clark, A. Being there; why implementation matters to cognitive science. *AI Review 1*, 4: 231-244, 1987.

Clark, A., and Chalmers, D. The Extended Mind. Philosophy –Neuroscience – Psychology Research Report, Washington University, St. Louis, 1995.

Clark, A. *Being There: Putting Brain, Body, and World Together Again*. MIT Press, 1997.

Husserl, E. *Ideas: General Introduction to a Pure Phenomenology*. Trans. W.R. Boyce Gibson. Allen and Unwin, 1931.

Husserl, E. *Cartesian Meditations: An Introduction to Phenomenology*. Trans. Dorian Cairns. Martinus Nijhoff, 1960.

Lipson, H., and Pollack, J.B. Automatic design and manufacture of robotic lifeforms. Nature, 406: 974-978, 2000.

Maturana, H.R. The origin and conservation of self-consciousness. Kybernetes, 34(1/2): 54-58, 2005.

Maturana, H.R., and Varela, F.J. Autopoiesis and cognition- The realization of the living. D. Reidel Publishing, 1980.

Maturana, H.R., and Varela. F.J., *The Tree of Knowledge: The Biological Roots of Human Understanding.* New Science Library, 1987.

Merleau-Ponty, M. *Phenomenology of Perception*. Trans. Colin Smith. Routledge and Kegan Paul, 1962.

Merleau-Ponty, M. *The Structure of Behavior*. Trans. Alden Fisher. Beacon Press, 1963.

Merleau-Ponty, M. Eye and mind. *The Primacy of Perception and Other Essays*. ed. James M. Edie. Northwestern University Press, 1964.

Pfeifer, R., and Scheier, C. *Understanding Intelligence*. MIT Press, 1999.

Prinz, J. Level-Headed Mysterianism and Artificial Experience. *Journal of Consciousness Studies*. 10: 111-132, 2003.

Sharkey, N.E., and Ziemke, T. Life, Mind and Robots- The Ins and Outs of Embodied Cognition. *Hybrid Neural Systems*. eds. S. Wermeter and R.Sun. Springer Verlag, 2000.

Thompson, E. Life and mind: From autopoiesis to neurophenomenology- a tribute to Francisco Varela. *Phenomenology and the Cognitive Sciences*. 3: 381-398, 2004.

Varela, F.J., Thompson, E., Rosch, E. *The Embodied Mind*. MIT Press, 1993.

Weber, A., and Varela, F.J. Life after Kant: Natural purposes and the autopoietic foundations of biological individuality. *Phenomenology and the Cognitive Sciences*. 1(2): 97-125, 2002.

Ziemke, T. Are Robots Embodied? Paper presented at the first International Workshop on Epigenetic Robotics: Modeling Cognitive Developments in Robotic Systems, Lund, Sweden, 2001.

Ziemke, T. What's that thing called embodiment? *Proceedings of the 25th Annual Meeting of the Cognitive Science Society*. Lawrence Erlbaum, 2003.

Towards Streams of Consciousness; Implementing Inner Speech

Pentti O A Haikonen
Nokia Research Center
P.O. Box 407, FI-00045 NOKIA GROUP
pentti.haikonen@nokia.com

Abstract

Inner speech is an aspect of human cognition that has been largely neglected by traditional artificial intelligence research. It is argued here that inner speech is an important contributor to cognition and consciousness and therefore also conscious machines should incorporate it. The realization of inner speech in machines involves also notoriously difficult linguistic issues, like sentence understanding. Here an approach to language processing by associative neural networks is proposed as the solution. This method works without explicit parsing or grammatical rules. The cognitive effects of inner speech arise from its content; inner speech is about something and that content affects the operation and behavior of the cognitive system. Consciousness involves the awareness of the mental content; inner speech is seen here as one tool for introspection that facilitates this awareness. In inner speech we may comment ourselves in a way that we have learned from others. This self-appraisal is seen as a process that leads to enhanced social self-awareness and self-image.

1 Introduction

In humans the inner or silent speech is the "little voice inside the head" that commences when we awake and ceases when we fall asleep. Inner speech seems to be present also in dreams at least to some degree. Inner speech is persistent; it is difficult to suppress it for any extended moment while awake. In folk psychology inner speech is often equated to thinking and is understood as a main difference between man, animals and machines. Introspection may mislead us, but inner speech would seem to be one consciousness-related phenomenon that we can be rather sure of. Inner speech is a tool of introspection; via the flow of inner speech we are able to report to ourselves what we think. When we fall asleep the flow of inner speech stops and our consciousness is very much diminished. Nevertheless, it is obviously possible to be conscious to at least some degree without language, solely by the flow of sensory percepts, inner imagery, feelings, actions, needs and the like.

Inner speech has been traditionally ignored by AI researchers while within cognitive psychology and neuroscience its potential as a key component of consciousness has been seen (for instance Morin & Everett 1990, Morin 1993, 1999, 2003, 2005, Siegrist 1995, Schneider 2002). Lately however, also some machine cognition researchers have recognized the importance of inner speech. (Clowes & Morse 2005, Haikonen 1998, 1999, 2000, 2003, 2005a, 2005b, Steels 2003a, 2003b). Also Duch (2005) has proposed a conscious architecture with a flow of "mind objects"; words and images.

In the context of machine consciousness inner speech has a rather crucial position as its explanation and artificial generation involves almost every other issue of cognition; perception, recognition, the grounding of meaning, situational inner models, the temporal handling of information; what the situation is now, what it was before, what has changed, etc. It seems obvious that a machine cannot have meaningful inner speech if it does not understand the world, as this would be a prerequisite for the understanding of language. The solving of the issues of inner speech would involve the solving of most of the practical problems of conscious machines.

The author sees the engineering challenges of inner speech as two-fold. The first issue relates to the enabling neural mechanisms and supporting circuitry for inner speech. The second issue relates to the contents of inner speech, how its meaning is grounded, how it arises and what are its effects on cognition and consciousness, especially self-awareness and self-image. In the following the neural and linguistic prerequisites are treated first and the consciousness-related issues next.

2 Mechanisms for Inner Speech

Natural language understanding is a hard problem that has not yet been solved satisfactorily and definitely not in any elegant way. Yet this is the exact problem that must be solved if meaningful inner speech is to be created in a machine. The author's "multimodal model of language" (Haikonen 2003) is one attempt towards natural use and understanding of language in a machine. Here an experiment relating to the implementation of this approach with associative neural networks is described.

Spoken words are temporal sound patterns consisting of sequences of phonemes. The detection of words calls for the ability to capture and analyze sound patterns and transform the serial phoneme sequence into a parallel representation. Thereafter there are two possibilities for the word representation, namely the distributed representation and the single signal (grandmother) representation. In the distributed representation there can be one or more signals per phoneme or syllable, thus each word will be represented by a signal vector. In the single signal representation each word is represented by one signal only. The distributed representation method is more flexible and allows the use of inflection while the single signal method is easier to use in simple simulations.

The author has used an associative neuron group (Haikonen 1999) as the basic processing unit for the distributed and single signal representations. The operation of the associative neuron group is explained here in simplified (but working) terms, which can be readily implemented with a computer program. The associative neuron group can be seen as a group of neurons that share common associative (synaptic) input signals. Thus their synapses form a kind of a matrix, figure 1.

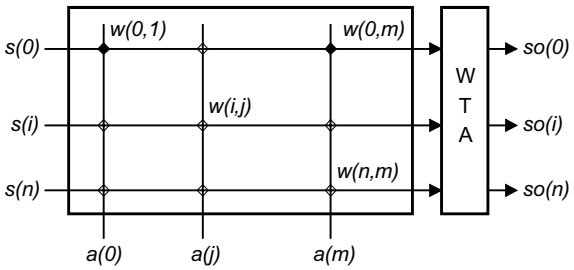

Figure 1. The associative neuron group

Each cross-point can be understood as one synapse and each horizontal line can be understood as one neuron with *m* synapses and one output signal *so(i)*. The purpose of each synapse is to associate the crossing signals *s(i)* and *a(j)* with each other. This is done via the synaptic weight *w(i,j)*. The synaptic weight value *w(i,j)* = 0 means that the signals *s(i)* and *a(j)* are not associated with each other, while the value *w(i,j)* = 1 means that the signals *s(i)* and *a(j)* are associated with each other.

The associative link between the two signals *s(i)* and *a(j)* is created if they appear simultaneously. The synaptic weight value *w(i,j)* is computed as follows at the moment of association:

(1) $w(i,j) = s(i)*a(j)$

where

s(i) = the input of the associative matrix; zero or one
a(j)= the associative input of the associative neuron group; zero or one.

Initially the synaptic weight value *w(i,j)* has the value of zero. The synaptic weight value *w(i,j)* = 1 gained at any moment of association will remain permanent. In the figures the symbol ♦ at the line crossings is used to indicate a synapse with the weight value 1. (A correlative learning rule for more general learning is given in Haikonen 1999, also described in Haikonen 2003, p. 78)

The associated signal *so(i)* is evoked by the signal *a(j)* according to (2) and (3). First, for each *so(i)* signal an evocation sum *Σ(i)* is computed as follows:

(2) $\Sigma(i) = \Sigma\, w(i,j)*a(j)$

where

Σ(i) = evocation sum
w(i,j) = synaptic weight value; zero or one.

Next, the output *so(i)* is determined by using an output threshold that equals to the maximum evocation sum. This method is also known as the Winner-Takes-All threshold (WTA).

(3) $so(i) = 0$ *IF* $\Sigma(i) < threshold$
$so(i) = 1$ *IF* $\Sigma(i) \geq threshold$

where

$threshold = max\{ \Sigma(i)\}$

The state of the complete associative neuron group can be computed by the above equations by running the indexes from zero to *n* and *m*.

The associative neuron group can be applied to language processing neural networks as will be shown by the next example.

According to the "multimodal model of language" sensory modalities consist of feedback loops that are associatively connected and in this way try to broadcast their percepts to each other. The percepts are signal vectors where each individual signal represents a detected elementary feature. These elementary features are extracted from sensory infor-

mation via sensor-specific preprocesses. This perception process is also affected by the feedback from the system. (The author proposes that this kind of a system is conscious of an entity, when each sensory modality percept is about the same entity and represent different aspects of that entity, hence broadcasts are globally accepted and the whole system is in a kind of multiple closed-loop state.)

Thus, according to this model the language processing takes place in the auditory sensory modality, but is assisted by all the other modalities as well. The general outline of the auditory sensory modality with connections to elsewhere is depicted in the figure 2.

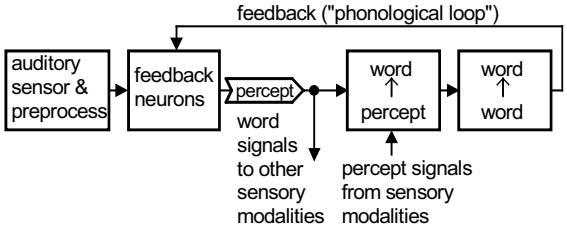

Figure 2. Linguistic modality model, part of the auditory sensory modality

The basic meanings of the words are grounded to sensory percepts like objects, sensations and change. These percepts are associated with words so that each percept may evoke the corresponding word (the percept→ word box in the figure 2). However, *our inner speech is not a list of the names of seen objects*, instead it is more like a running commentary about the perceived situation. Names are not important, possibilities, affordances (Gibson 1966) are. Here also, it should be seen that a perceived entity would evoke many kinds of responses in the other sensory and motor modalities; these would be perceived by those modalities and broadcast to the linguistic modality. Hence the evoked words would be related to the initially perceived object in a more general way. Also, the visual sensory modality is not the only relevant modality here; inner speech may be cued by other sensory modalities as well, like the auditory, touch, temperature, hunger. (Name→ percept association is important whenever a verbal description of a situation is to be transformed into a mental image of the same.) Nevertheless, the percept→ word association is a rather straightforward process and is not elaborated here.

The understanding of a sentence calls for the ability to extract the relationships between the entities that are described by the words in that sentence. There is also a syntactic component; part of the meaning is encoded in the word order and/or in the inflection of the words. This process would be executed in the word→ word association box in the figure 2.

A more complicated associative neural network is required for this word→ word association process. A simple example is presented here in order to illuminate the relevant basic issues and requirements.

In this example each word is represented by one dedicated signal (single signal representation). Distributed representation would also have been possible as was already done by the author (Haikonen 1999).

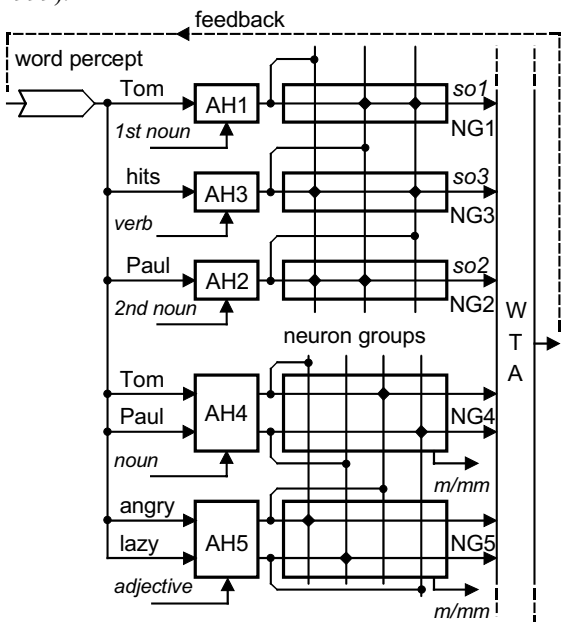

Figure 3. An associative network for linguistic word→ word processing

The associative neural network of the figure 3 is able to process simple sentences like those that describe action between two persons. Furthermore, these persons may or may not be characterized by adjectives. The network consists of neuron groups NG1, NG2, NG3, NG4 and NG5 which all share a common WTA-threshold circuit. The circuits AH1 ... AH5 are "Accept-and-Hold" circuits that recognize individual signals and hold them for a short period. The operation of the "Accept-and-Hold" circuits is grounded; each circuit accepts only its own kind of words, nouns, verbs and adjectives. This is facilitated here by defining each word as a noun, verb or adjective when the vocabulary is taught. In real robotic applications this process would take place during learning of words in natural environment.

Also the position of the word within the sentence matters; first and second nouns are captured separately. The figure 3 is simplified; only those signal lines that are relevant to the specific example are shown.

The subject-object action is captured from the incoming sentence by the neuron groups NG1, NG2

and NG3 and the related Accept-and-Hold circuits AH1, AH2 and AH3. The first noun and the second noun Accept-and-Hold circuits AH1 and AH2 are connected. They accept nouns sequentially; AH1 circuit accepts and captures the first noun and AH2 circuit captures the second noun. AH3 circuit accepts the verb.

When the network learns the information content of a sentence it forms associative connections between the words of the said sentence. The sentence as such is not stored anywhere in the network. As an example the sentence "*Angry Tom hits lazy Paul*" is considered. During learning, that is, when the network receives the sentence, certain associative links are formed. These links operate via synaptic weight values of 1 and are depicted in the figure 3.

The understanding of a sentence involves the ability to answer to questions about the information content of the sentence. When, for instance, the question "*Who hits Paul*" is entered, the word "who" is captured by the AH1 forcing the word "Paul" to be captured by AH2. The verb "hits" is captured by AH3. The associative connections will give the correct response "Tom". The question "*Paul hits whom*" will not evoke incorrect responses, as "Paul" will be captured by AH1 and in that position does not have any associative connections.

The associative neuron groups NG4 and NG5 associate nouns with their adjacent adjectives. Thus "Tom" is associated with the adjective "angry" and "Paul" with "lazy". This is done in the run; as soon as "Tom" is associated with "angry", the Accept-and-Hold circuits AH4 and AH5 must clear and be ready to accept new adjective-noun pairs. After successful association the question "Who is *lazy*" will evoke the response "Paul" and the question "Who is *angry*" will evoke the response "Tom".

Interesting things happen when the question "Is *Tom lazy*" is entered. The word "Tom" will evoke the adjective "angry" at the output of NG5 while the word "lazy" will evoke the word "Paul" at the output of NG4. Both neuron groups NG4 and NG5 have now mismatch condition; the associatively evoked output does not match the input. The generated match/mismatch signals may be associated with the words like "yes" and "no" and thus the system may be made to answer "No" to the question "Is *Tom lazy*" and "Yes" to the question "Is *Tom angry*".

This exercise was executed in the form of a Visual Basic program. The visual interface of this program is shown in the figure 4. This picture presents the situation when the example sentence and some questions about the information content of the sentence have been entered.

This exercise shows that it is possible to use associative neuron groups for language processing, at least for simple sentences. The importance of the grounding of meaning is also demonstrated; while the actual meanings of the words are not grounded here, the categorical meanings are and this grounding is still essential. Further refinement of this approach would involve the basic grounding of meaning for the words and additionally, the use of *inner situational models* that would involve representations in other sensory modalities. These models would allow the grounding and inspection of the relationships between the entities of a given sentence and would also facilitate paraphrasing.

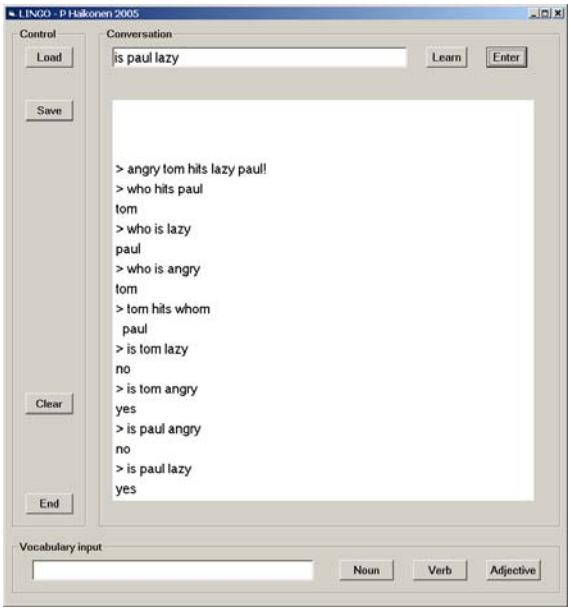

Figure 4. Sentence understanding with the associative neural architecture, a Visual Basic program

A complete cognitive system with the flow of inner speech as sketched by the author (Haikonen 2003) would use these kinds of neural systems as subsystems within the auditory modality. It is worth noticing the simplicity of this approach; this kind of linguistic processing does not utilize explicit sentence parsing or grammatical rules. There is no innate grammar, the "grammar" of a sentence arises from the relationships of the real world.

3 Inner Speech and Consciousness

As good as the devised neural machinery might be it would only be a supporting platform. Any phenomena that relate to consciousness would arise via the content that were carried by the platform in the form of inner speech. After all, our inner speech is about something and, on the other hand, we are not able to perceive the physical neurons or neural processes as such behind our inner speech. The contents of inner speech would allow us to shape our aware-

ness while other, transparent neuron and architecture-related mechanisms, especially feedback and cross-association, would allow the content to be introspected and controlled by itself.

Would it be possible for one system to introspect and control itself? Even simple feedback control systems can do this. However, in this case there may be two major systems interacting with each other, namely the auditory-linguistic modality and the speech-motor modality. These modalities would usually carry the same information albeit in different terms; "heard" word representations and motor command representations for spoken or overt speech. This arrangement would allow the easy inspection of inner speech as a copy of it would be available in the speech-motor modality. Ryding et al (1996) propose: "Audible and silent speech may represent two principally different types of cerebral feedback systems, one for overt sensory–motor activity and one for a pure internal cognitive feedback". There is also some experimental proof that inner "heard" speech and overt speech have separate neural substrates. For instance aphasic patients may complain that the words they speak are not the words they think and intend to say (Huang et al 2001).

The author has proposed that consciousness arises in a multimodal system from associative interconnections between the modalities (Haikonen 2003). According to the "multimodal model of language" inner speech would be one manifestation of these interconnections. If these interconnections break down, then inner speech and consciousness should also vanish. Indeed, Massimi et al. (2005) have noticed that during sleep, when there is no consciousness (or inner speech), neural communication between different parts of the cerebral cortex breaks down while local activities may still exist.

In inner speech we may engage in thoughts about thoughts: "I am thinking now" and in doing so be aware of having thoughts. Obviously this observation of one's own thoughts and the recognition of the ownership of the same would seem to be one manifestation of self-consciousness.

Duval and Wicklund (1972) define self-awareness as the state of being the object of one's own attention. This would include the paying of attention to one's own mental content such as percepts, thoughts, emotions, sensations, etc. Inner speech has been seen as a tool for introspection and one of the most important cognitive processes involved in the acquisition of information about the self and the creation of self-awareness (Morin 1990, 2005, Haikonen 2003, pp. 256 – 260).

With inner speech one can comment one's own situation. Morin (2005) sees this self-talk as a device that can reproduce and extend social mechanisms leading to social self-awareness. (The author has argued elsewhere that basic self-awareness does not require social interaction, see the "hammer test" in Haikonen 2003 p. 161.) As a part of social interactions we are subject to comments about ourselves, the way we are and behave. Self-talk allows us also to internally imitate the act of appraisal; we may echo the patterns of others' comments directly as such or as first-person transformations. We may ask ourselves: "Why did *you* do this stupid thing?" or "Why did *I* do this stupid thing?". Originally it was your mother that posed the question (Haikonen 2003 p. 240). In this way inner speech turns into a tool for self-evaluation, which in turn will affect our self-image; who we are, what we want.

4 Conclusion

Inner speech has been largely neglected by traditional artificial intelligence research perhaps because the algorithmic solving of problems in binary computers does not necessitate it. However, cognitive machines would be different. The emulation of the processes of the human brain and mind would be incomplete without the realization of inner speech.

Unfortunately the realization of inner speech in machines involves also notoriously difficult linguistic issues, like the grounding of meaning and sentence understanding. Here an associative neural approach that works without explicit parsing or grammatical rules is outlined and verified to a limited degree by a computer simulation program.

Inner speech is about something and that content affects the operation and behavior of the cognitive system. Consciousness involves the awareness of the mental content; conscious beings may introspect their mind. Inner speech is seen here as one tool for introspection that facilitates this awareness. Inner speech is not only a running commentary of external events, it involves also self-appraisal. This self-appraisal is seen as a process that leads to enhanced social self-awareness and self-image.

For practical reasons robots should have inner speech, as this would allow communication with natural language in natural way. This would allow easy peeking into the workings of the robot brain; technically it would be very easy to monitor and listen to the inner speech. Also, from a philosophical point of view it would be easier to accept that a robot thinks if it had the flow of inner speech and imagery in a similar way that we have.

Inner speech helps us to make sense of our moment-to-moment existence. A conscious robot should experience its existence in the same way. Therefore we should build machines with inner speech, machines that have streams of consciousness.

References

Clowes, R., Morse, A. F. (2005). *Scaffolding Cognition with Words*. Retrieved on 16. 12. 2005 from http://www.cogs.susx.ac.uk/users/robertc/Papers/ScaffoldingCognitionWithWords.pdf

Duch, W. (2005). Brain-Inspired Conscious Computing Architecture. *The Journal of Mind and Behavior* Vol. 26 (1-2) 2005, pp. 1 - 22

Duval, S., Wicklund, R. A. (1972). A theory of objective self awareness. New York: Academic Press

Gibson, J.J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Haikonen P. O. (1998). Machine Cognition via Associative Neural Networks. *Proceedings of EANN'98* pp. 350 – 357

Haikonen, P. O. (1999). *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

Haikonen, P. O. (2000). An Artificial Mind via Cognitive Modular Neural Architecture. *Proceedings of the AISB'00 Symposium on how to design a functioning mind* pp. 85 – 92. UK: University of Birmingham.

Haikonen, P. O. (2003). *The Cognitive Approach to Conscious Machines*. UK: Imprint Academic.

Haikonen, P. O. (2005a). Artificial Minds and Conscious Machines. In D. N. Davis (Ed.) *Visions of Mind: Architectures for Cognition and Affect* pp. 286 – 306. USA: Idea Group Inc.

Haikonen, P. O. (2005b). You Only Live Twice; Imagination in Conscious Machines. In R. Chrisley, R. W. Clowes & S. Torrance (Eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.* The Society for the study of Artificial Intelligence and the simulation of behaviour, UK. pp. 19 – 25.

Huang, J., Carr, T. H., Cao, Y. (2001). Comparing Cortical Activations for Silent and Overt Speech Using Event-Related fMRI. In *Human Brain Mapping* 15 (2001), pp. 39 – 53.

Massimi, M. et al. (2005). Breakdown of Cortical Effective Connectivity During Sleep. *Science*, Vol. 309 30 Sept. 2005 pp. 2228-2232

Morin, A., Everett, J. (1990). Inner speech as a mediator of self-awareness, self-consciousness, and self-knowledge: an hypothesis. *New Ideas in Psychol*. Vol 8. 1990, No. 3, pp. 337 - 356

Morin, A. (1993). Self-talk and self-awareness: On the nature of the relation. *The Journal of Mind and Behavior*, 14. pp. 223-234.

Morin, A. (1999). On a relation between self-awareness and inner speech: Additional evidence from brain studies. *Dynamical Psychology: An Interdisciplinary Journal of Complex Mental Processes*. Retrieved from http://cogprints.org/2557/ on 14.12.2005.

Morin, A. (2003). Let's Face It. A review of The Face in the Mirror: The Search for the Origins of Consciousness by Julian Paul Keenan with Gordon C. Gallup Jr. and Dean Falk. *Evolutionary Psychology*, 1:161-171.

Morin, A. (2005). Possible links between self-awareness and inner speech: Theoretical background, underlying mechanisms and empirical evidence. *Journal of Consciousness Studies*. Volume 12, No. 4-5, April-May 2005

Ryding, E., BraÅdvik, B., Ingvar, D. H. (1996). Silent Speech Activates Prefrontal Cortical Regions Asymmetrically, as Well as Speech-Related Areas in the Dominant Hemisphere. *Brain and Language* Volume 52, Issue 3 (March 1996), pp. 435-451

Siegrist, M. (1995). Inner speech as a cognitive process mediating self-consciousness and inhibiting self-deception. *Psychological Reports,* 76, pp. 259-265

Schneider, J. F. (2002). Relations among self-talk, self-consciousness, and self-knowledge. *Psychological Reports*, 91: 807-812.

Steels, L. (2003a). Language Re-Entrance and the "Inner Voice". In O. Holland (Ed.), *Machine Consciousness*, pp. 173 – 185, UK: Imprint Academic

Steels, L. (2003b). Evolving grounded communication for robots, *Trends in Cognitive Science*, 7(7), July 2003, pp. 308 – 312.

Could a Robot have a Subjective Point of View?

Dr Julian Kiverstein
Dept of Philosophy, University of Edinburgh
3rd Floor, David Hume Tower,
George Square, Edinburgh, EH8 9JX
j.d.kiverstein@sms.ed.ac.uk

Abstract

An argument for the possibility of conscious robots would have to show that the brain is neither necessary nor sufficient for the possession of consciousness. I will set about giving just such an argument. Proponents of the enactive theory of perception have argued that neural activity doesn't always suffice for the having of conscious experience. They have argued that the body and environment can also play a constitutive role in enabling conscious experience.

In this paper I will argue for the stronger claim that neural activity isn't necessary for conscious experience either. A robot could, I will argue, enjoy phenomenal consciousness. This has been denied by at least one prominent proponent of the enactive theory of perception (see Alva Noë (2005. 230)) who has argued that a robot wouldn't count as a subject of experience. In the absence of a subject of experience, Noë thinks it makes no sense to attribute phenomenal consciousness.

I will argue that on the contrary a robot could be a subject of experience. My argument will proceed in three stages. The first stage argues that a creature is a subject of experience if it has a first-person perspective. I set out some conditions a creature must satisfy if we are to attribute to that creature a first-person perspective. The most important of these conditions is that the representations the creature produces must have reflexive content – they must, in a sense I explain, be representations that refer to themselves.

The second stage of my argument uses a variation on Andy Clark's (2000) argument for the conclusion that access implies qualia. I claim that any representation that has reflexive content will be one to which we have access. Clark has argued that access implies qualia, so it follows that a representation with reflexive content will also have qualia.

The final step in my argument will be to show that action-oriented representations (see Clark 1997 for an account of this type of representation) have reflexive content. Many robots that are capable of producing adaptive behaviour do so by means of action-oriented representations. These robots, I will argue, already meet the conditions for having a first-person perspective. Thus robots with a low-degree of phenomenal consciousness I will claim already exist.

My paper will finish by attempting to motivate this conclusion through a reflection on the connection between consciousness and life. Robots that produce adaptive behaviour are models of life. I will argue that because of the connection between consciousness and life these robots are also models of consciousness.

References

Clark, A. 2000: ‘A case where access implies qualia’ in *Analysis* 60.1: 30-8

Clark, A. 1997: *Being-There: Putting Brain, Body and World Together Again* (Cambridge, MA: MIT Press)

Noë, A, 2005: *Action in Perception* (Cambridge, MA: MIT Press)

Acting and Being Aware

Jacques Penders
Sheffield Hallam University
Sheffield
j.penders@shu.ac.uk

Abstract

One often assumes that we, rational human beings, first think and than act. This paper is an attempt to describe the mental characteristics governing the performance of regular everyday actions; and shows that no mental act has to precede our actions, instead of consciously thinking before we act, we mostly act while simultaneously overseeing our acting. The case of ball juggling is used to underpin the analysis with practical facts.

1 Introduction

In the overview paper of the 2005 Machine Consciousness conference the goals of Machine Consciousness are described as: 1) to create artifacts that have mental characteristics typically associated with consciousness (such as awareness, self-awareness, emotion and affect, experience, phenomenal states, imagination etc.); and 2) to model these aspects of natural systems in embodied models (e.g., robots), (Chrisley et al., 2005).

This definition stipulates that the mental phenomena are to be studied in an embodied creature or model, thus the combination of mental states and physical action is brought into the focus. The theme of the current conference concerns *"models which show the emergence of, or otherwise treat, processes or systems underlying these core themes."* The present paper addresses this theme with an attempt to unravel the mental characteristics, which manifest themselves in regular action oriented contexts. I try to describe the mental stance applied by a human being while performing the standard routines of everyday life. Without being able to systematically order all the mental characteristics mentioned above I will discuss a few assumptions so as to indicate some ordering and suggest a place for the stance I am describing.

An often-encountered assumption – which I believe is generally untrue - is that a certain mental act precedes our bodily actions, or in plain language that we first think and then act. For instance Haggard et al. (2002) write: "*Normal human experience consists of a coherent stream of sensorimotor events, in which we formulate intentions to act and then move our bodies to produce a desired effect*".

However, William James (1890) already clearly noted that the suggested ordering in time does not hold. He described his concept of ideomotor action summarised as: we think the act and it is done. An example of his: "*We think to drink our coffee and we find ourselves already holding the cup in our hands*".

I will argue a step beyond and show that we often act before any conscious thinking has occured. My point is not to substantiate a general moral excuse for cases where we have done things, which we afterwards regret. My point is pragmatic: we cannot act and behave as we do in ordinary life if we first have to think (let alone think over) every action. Being human, we like to think of ourselves as rational beings. In the history of Philosophy Immanuel Kant is probably the clearest exponent of this view. He saw a human being as a logical subject of thought (Stuart, 2005) that is bound to act in the physical world. Kant's work could be seen as a major attempt to reconcile the two while giving primacy to rationality. And indeed on occasions we do first think and then try to act accordingly. However, considering the full extent of all the actions an individual performs in his or her everyday routine, it is clear that our rationality can operate only in the background. The occasions where thinking precedes acting are the exception and not routine practice.

In the morning of a regular day, while deliberating on how to make the best out of the day of today, we routinely drink our coffee and make our way to work, say by car. While driving the car, we suddenly stand on the brakes as we are forced to an emergency stop. Only after having come to a stand-

still we come to think about what we have done the seconds before.

Instead of first thinking and then acting, we only oversee our actions with our conscious and rational minds. I call the mental stance which we take when driving the car and which generally prevails when we act: **being aware without focus.** Interesting about this stance is that actions are selected and performed without them being in the focus of attention, and what is more, as I will show below, when attention gets focussed it often interrupts the actions. I use juggling as an example to investigate the flow of the mental processes.

It is interesting on its own to unravel the mental stance in which action selection takes place, since it might shed light on the complex of mental states and stances by which a human being monitors and controls his of her body and actions. Definitely the human body on its own is a complex system with a complex control structure, the understanding of which could function as a paradigm for robot and machine design.

2 Attention and Acting

In order to explain the stance of being aware without focus, first a few words about the closely related notion of attention. Our mind can be in different modes of activity, with sleeping as the extreme on one end. When awakening from sleep, our mind has to "warm-up" in an arousal phase. Then we become generally aware enough so that we can attend: the mind is aroused and proceeds via getting aware to attention. Further onwards, when there is attention, consciousness and conscious experiences may come in.

Attention is since Broadbent's work often conceived of as a filter for or a gate to consciousness, which blocks, weakens or inhibits incoming messages from the senses. Baars (1997) introduced the metaphor of attention acting as a spotlight in a theatre. When in the spotlight of attention, the mental processing becomes accessible to consciousness. The filter metaphor characterises the operations of attention as reductive while the spotlight metaphor suggests amplification; both nevertheless agree that attention is selective.

Attention also has to do with action. "Awareness [or being aware] implies perception, a purely sensate phase of receptivity. Attention reaches. It is awareness stretched toward something. It has executive, motoric implications. We attend **to** things." (Austin, 1998).

Appropriate applications of motor skills - that is to act appropriately - requires a proper combination of perception, action selection and action execution. The role of attention in relation to perception has been widely studied; however its role in applying motor-skills has not received as much scientific interest. The reason for this might be that motor-control, which is a prerequisite for motor-skilfulness, is very much on and below the edge of what we can consciously experience and control.

The performing arts and sports sciences deal with action and attention. Artists and sporting men and women engage in what is called *deliberate practice* (Rossano, 2003) (Ericsson et al., 1993): the concentrated effort to hone and improve specific (mental and) physical skills. Literature on deliberate practice distinguishes between external attentional focus and internal attentional focus; internal attentional focus means that the performer directs attention to the movements itself, while in external attentional focus, the attention goes to the effects the movements have on the environment (Wulf and Prinz, 2001). In both attitudes attention plays a prominent role, and generally external attentional focus is more proficient.

The influence of internal attentional focus may be observed in for instance dancing or martial arts classes. In a class of beginners, the students might be quite able to straightforwardly copy the movements of their instructors. However, when the instructor explains the consecutive moves to the very detail, several students appear not to be able to perform, even though they did so before. And reverse, when the instructor is asked about the details of a move which (s)he has never made explicit before, it is likely he or she has to perform first before being able to explain. Applying attentional and conscious control in motor-control hampers performance. Extreme examples are observed with patients suffering from the syndrome called apraxia. Apraxia denotes the inability of a patient to perform a certain skilled movement. For instance when asked to demonstrate teeth brushing, the patient is unable to do so, whereas he or she is perfectly able to brush the teeth in the morning.

Attention obviously has motoric implications, the examples show that internal attentional focus and conscious control of motor-skils may even lead to an inability to act.

The notion of external attentional focus, is not clearly defined and allows several interpretations. In a narrow, but easiest to define sense it denotes attention focusing on bringing about a single effect: directing a tennis ball, or throwing a single ball or bean bag into the air such that it can be caught. I will test this reading in the next section in the context of juggling.

3 Acting and Awareness

Five-ball juggling is hard and requires fast acting, the complication being that between throwing and catching the same ball four other objects – three of which are already up in the air - have to be handled. When first starting, it is a problem to throw each of the five balls one after the other before the first has returned (flashing as it is called), in doing so a novice will not be able to tell which ball was first thrown, let alone be able to catch it with the proper hand.

The novice juggler is trying to apply full and conscious attention, and that leads him or her astray. In juggling, the time lapse between throwing and catching a single ball is not more then a single second. Meanwhile, in five ball juggling four other objects are flying around appealing for attention. However, it is known that per second no more then two attentional shifts can occur, which is far too slow for five-ball juggling.

Juggling combines perception with action; in the one second between throwing and catching a particular ball four other objects have to be handled as well. Psychological experimentation has shown that the time required for the single voluntary act of pressing a button *only* when a light flashes is about 0.15 seconds (Austin, 1998). In contrast, observations of jugglers show that the time lapse between two catches of the same hand may be as little as 0.2 seconds (Polster, 2003). In this short interval several actions of this hand flow into each other: catching, bringing to throwing position (dwelling), throwing and preparing/waiting for the next, while in the middle of this series the other hand has to start its own series as well; refer to Polster (2003) for more details. A simple comparison of the time required for a voluntary act and the constraints of juggling shows the impossibility of juggling being a series of voluntary actions.

Because of the complexity and time constraints in five-ball juggling, correction of the movements and abandoning systematic flaws is quite difficult and requires persistence and endurance. An explanation is that there exist two independent systems or circuitries for the perceptual control of movement (Rossano, 2003). Raichle (1997) makes a distinction between "the neural circuitry underlying the unpractised, presumably conscious performance of a task on the one hand, and the practised presumably non-conscious performance of a task on the other hand." The response time of the latter circuitry is significantly shorter than that of the first (Raichle, 1997).

Voluntary actions are slow compared to involuntary acts, for instance a reflexive jerk takes only 0.025-0.05 seconds, which is in the order of five times faster then a voluntary act!

Internal attentional focus hampers execution of actions and actions are generally slower then when external attentional focus is applied. In five-ball juggling external attentional focus fails as there is not enough time to focus attention. Obviously the very fast, but complex and precision requiring moves in juggling cannot be under full conscious control. The juggler must be applying a different stance: a very sensate stance requiring awareness but avoiding any attentional focus; I call this stance: **being aware without focus.**

Indeed, an experienced juggler does not focus on the individual balls. In his juggling book Dancey (1994) advises: "*While learning [a five-ball pattern] you are trying to* make *yourself do it, when you can do it you* watch *yourself doing it."*

In five-ball juggling, there simply is not enough time to focus attention; restricting attention results in faster actions. However the surprising thing is that when no full attention is required for acting, the mind performs other tasks concurrently.

In daily life we perform many actions without attentional focus, for instance when walking the body performs an intricate combination of muscle activities to maintain posture; car driving and juggling are other examples. Three-ball juggling is less demanding than five-ball juggling. While juggling, the juggler can do other things as well, for instance speak, walk etc.; however non-focussed awareness is permanently required, when the juggler's attention drifts away and focuses elsewhere the balls drop. Car driving implies a similar requirement; the driver can perform many other things while driving but a certain level of awareness is required throughout.

I have avoided any attempt to define the notion of attention; therefore I cannot conclude that attention is not involved in the stance of being aware without focus. But referring to the spotlight metaphor, if there is attention involved, it is only a dim light. Because attention is a preliminary for consciousness this conclusion has implications for the role of consciousness as well.

The juggling example shows that no conscious mental act is required in order to perform, and what is more it shows that for fast acting no conscious mental act **can** precede the execution of the actions.

4 Acting and Emotions

Many cognitive scientists subscribe to the view that affect addresses the problems of decision making and action selection (Shanahan, 2005; Sloman, 2001). However, in the state of being aware without focus, the influence of affect seems much reduced.

Returning to the example of routinely driving the car on the way to work; our conscious mind was occupied of our plans for the coming day, and we

were at a sudden interrupted by the emergency break. The action of pressing the breaks was a straight reaction to events occurring around us, and as far as I can see it was not guided by any obvious emotion. Of course, emotions come up afterwards and may interfere with our consciously reconstructing the events, but they did not initiate nor guide the breaking action.

Literature on deliberate practice refers to emotions mainly by advising to attain an optimal emotion state and thinking positively (Wulf and Prinz, 2001).

Some descriptive evidence about the interference of emotions with acting can be found in the area of the eastern martial arts, in particular where Zen-Buddhism is involved. The aim of Zen-Buddhism is to voluntarily move into and try to intensify a mental state described as: "*When the ultimate perfection is attained, the body and limbs perform by themselves what is assigned to them to do with no interference from the mind. [The technical skill is so autonomised it is completely divorced from conscious efforts].*" (Takuan, translated in Suzuki 1959, the addition in brackets by Suzuki). The stance of being aware without focus, which I try to describe, bears similarities. Thus, though the aims are quite different, the Zen related literature contains interesting observations concerning the influence of affect and emotion on acting.

In Japanese, the state of perfection is called Mushin, which literally means "no-mind" or "without mind, without heart" (Austin 1998). Descriptions of this state are found in Hinduism as well; an interesting metaphor is given in the text called The Bhagavat Gita, it says that someone who masters this state, " … *withdraws all his senses from the attractions of their objects, even as a tortoise withdraws all its limbs*,…" (BG 2,58). The citation does not imply that the senses are withdrawn; the point is the mental stance with respect to the 'attractions' of the senses. Austin (1998) gives a further addition: "*The no mind of Zen implies a mental posture in which at least two things are going on: (1) bare attention still registers percepts, but (2) there are no emotional reverberations."*

The impact of emotion on performing is also described by the 20th century Zen master Taisen Deshimaru in a discourse for martial art practitioners: *"If our mind is upset, the natural functions of our bodies also tend to be disturbed. When the mind is calm, the body can act spontaneously ..."* (Taisen Deshimaru, 1982). In the ideal attitude of the swordsman this is pushed to the limit: "*The perfect swordsman takes no cognisance of the enemy's personality, no more than of his own. For he is an indifferent onlooker of the fatal drama of life and death in which he himself is the most active participant.*" (Suzuki, 1959).

Though my evidence on emotions is rather thin, I tend to conclude that intense emotions have a similar effect on performance and acting as focussed attention has.

Interesting to note at this point is an approach to deliberate practice developed Singer (1985, 1988) with aims at non-focused performance. Wulf and Prinz (2001) call it mysteriously *"a compromise between awareness and nonawareness strategies"*. It includes several steps: *readying* or arousal; *imaging* that is, going through the motion mentally; *focusing*, concentrating on a certain cue to block out all other thoughts; and *executing* the movement, while not thinking about the act itself or the possible outcome. This approach is much in line with the advices from Zen Buddhism, however it is seldom mentioned in the recent literature on deliberate practice.

4 Consciously Inhibiting Actions

A recent assumption in cognitive neuroscience is that the mind has a layered structure with at least three organising levels concerning body experience. "The lowest level is an assembly of neuronal information coming from all parts of the body; at the middle level the body schema are situated which secure the emergence of the conscious body image at the third level" (Yamadori, 1997). The body schema are subsystems 'implementing' James' ideomotor actions, for instance grabbing the coffee cup. Interesting for my analysis is the distinction between the second and the third level; are these levels really separate and may the second level operate independent from the third? The independence of the second level is shown by the split-brain studies and in particular very compellingly by the so-called *Anarchic hand* (Blakemore et al., 2002). The latter designates pathological behaviour in which a patient's right hand manipulates a tool properly but 'spontaneously', that is with the patient being aware of the hand acting, though neither consciously initiating the movement nor being able to inhibit the action. The anarchic hand shows that the neither attention nor consciousness are a prerequisite or a necessary condition (sine qua non) for action; they are not necessarily the initiator of actions. Moreover, it even shows that there exist pathological cases where consciousness is unable to inhibit actions.

Most people readily acknowledge that consciousness is not in control of the internal functioning of our body. The anarchic hand demonstrates that even skilful behaviour might be beyond the span of control of consciousness

Conclusions

I have made an attempt to describe the mental stance taken when performing regular everyday actions. I have called this stance *being aware without focus*; it is a stance in which there is typically little or no attentional focus.

Acting requires perception, action selection and action execution. These processes are often initiated and performed without any conscious deliberation; they are mostly on and below the edge of conscious experience and control.

Attention and emotions may interfere with acting but that often results in poorer or slower execution. Restricting attention results in faster actions. Surprisingly, if no full attention is applied for acting, the mind performs other tasks concurrently.

Attention is a gate to consciousness. Conscious thinking takes time and the often-supposed sequence that a mental act precedes bodily actions, or that we first think and then act cannot hold: it is too slow for many of our activities. In everyday practice we usually act before consciously thinking.

Conscious control is not a necessary condition for acting and consciousness only has weak control over the acting body, even though subjects have the feeling they consciously control their body.

Nevertheless, we do oversee our actions with our conscious and rational minds and except for pathological cases we are able to suppress many 'spontaneous' actions.

References

J.H. Austin, *Zen and the Brain*, MIT Press 1998.

B.J. Baars, *In the Theatre of Consciousness; The Workspace of the Mind,* Oxford University Press 1997.

S-J Blakemore, D.M. Wolpert and C.D. Frith, Abnormalities in the awareness of action, *TRENDS in Cognitive Sciences* Vol 6, no 6, 2002.

Chrisley, R., Clowes, R. W., & Torrance, S. "Next-generation approaches to machine consciousness". In R. Chrisley, R. W. Clowes & S. Torrance (eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment*, 2005.

C. Dancey, *Encyclopaedia of Ball Juggling*, Butterfingers, Bath UK 1994.

J. Decety, Do imagined and executed actions share the same neural substrate?. *Cognitive Brain Research*, 3:87-93, 1996.

Ericsson, K. A., R. Th. Krampe, and C. Tesch-Römer, 1993, 'The role of deliberate practice in the acquisition of expert performance.' *Psychological Review*, 100: 363-406.

Patrick Haggard, Sam Clark and Jeri Kalogeras. Voluntary action and conscious Awareness. *Nature Neuroscience* volume 5 no 4. 2002

W. James *The principles of Psychology*, 1890; Harvard University Press 1983.

B. Polster, *The mathematics of Juggling*, Springer-Verlag 2003.

M.E. Raichle, Automaticity: from reflective to reflexive information processing in the human brain, in: *Cognition, Computation and Consciousness*, K.Ito, Y. Miyashita and E. Rolls (eds), Oxford University Press, 1997.

M.J. Rossano, Expertise and the evolution of consciousness, *Cognition* Vol 89, (3) 2003

Shanahan, M. Consciousness, Emotion, and Imagination: A Brain-Inspired Architecture for Cogntive Robotics. In R. Chrisley, R. W. Clowes & S. Torrance (Eds.), Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment, 2005.

Singer, R. N. (1985). Sport performance: A five-step mental approach. *Journal of Physical Education & Recreation*, **57**, 82-84.

Singer, R. N. (1988). Strategies and metastrategies in learning and performing self-paced athletic skills. *Sport Psychologist*, **2**, 49-68.

Aaron Sloman, Beyond Shallow Models of Emotion. *Cognitive Processing* 2 (1), 177-198. 2003.

Susan Stuart, The Binding Problem: Induction, Integration and Imagination, ". In R. Chrisley, R. W. Clowes & S. Torrance (eds.), *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment*, 2005.

Taisen Deshimaru. *The Zen Way to the Martial Arts*, Arkana, Penguin Books 1982.

D.T. Suzuki, *Zen and Japanese Culture*, Princeton University Press 1959

Gabriele Wulf and Wofgang Prinz Directing attention to movement effects enhances learning: A review, *Psychonomic Bulletin & Review*, Volume 8, Number 4, 1 December 2001, pp. 648-660(13)

A. Yamadori, Body awareness and its disorders, in: *Cognition, Computation and Consciousness*, K.Ito, Y. Miyashita and E. Rolls (eds), Oxford University Press, 1997.

Using Emotions on Autonomous Agents. The Role of Happiness, Sadness and Fear.

Miguel Angel Salichs Maria Malfaz
RoboticsLab, Carlos III University of Madrid
28911 Leganés, Madrid, Spain
salichs@ing.uc3m.es mmalfaz@ing.uc3m.es

Abstract

This paper addresses the use of emotions on autonomous agents for behaviour-selection learning, focusing in the emotions fear, happiness and sadness. The control architecture is based in a motivational model, which performs homeostatic control of the internal state of the agent. The behaviour-selection is learned by the agent using a Q-learning algorithm while there is no interaction with other agents. In situations where interaction arises (e.g. interacting with other agents), agents rely on stochastic games approaches as a learning strategy. The agent is intrinsically motivated and his final goal is to maximize Happiness. The learning algorithms use happiness/sadness of the agent as positive/negative reinforcement signals. Fear is used to prevent the agent choosing dangerous actions or being in dangerous states where non-controlled exogenous events, produced by external objects or other agents, could danger him. Preliminary tests have been carried out in a virtual world, based in a role-playing game.

1 Introduction

The goal of our project is to develop social robots with a high degree of autonomy. The social aspect of the robot will be reflected in the fact that the human interaction will not be considered only as a complement of the rest of the robot's functionalities, but as one of the basic features.

For this kind of robots, the autonomy and emotions makes them to behave as if they were "alive". This feature would help people to think about these robots not as simple machines but as real companions. Evidently, a robot that has his own "personality" is much more attractive than one that simply executes the orders that he is programmed to do.

Emotions can act as a control and learning mechanism, driving behaviour and reflecting how the robot is affected by, and adapts to, different factors over time (Fong et al, 2002). In previous works (Malfaz and Salichs, 2004), an emotion-based architecture has been proposed.

Some researchers have also used emotions in robots. Most of them have made emphasis in the external expression of emotions (Breazeal, 2002) (Fujita, 2001) (Shibata et al 1999). Their robots include the possibility of showing emotions, by facial and sometimes body expressions. In this case, the emotions can be considered just as a particular type of information that is exchanged in the human-robot interaction process. In nature emotions have different purposes and interaction is only one of them. We intend to make use of emotions in robots trying to imitate their purpose in nature, which includes, but is not limited to, interaction. The role that plays each emotion and how the mechanisms associated to each one work are very specific. That means that each emotion must be incorporated to the robot in a particular way. In this paper we will present some basic ideas on how emotions such as happiness, sadness and fear can be used in an autonomous robot.

Emotions will be generated from the evaluation of the wellbeing of the robot. Happiness is produced because something good has happened, i.e. an increment of the wellbeing is produced. On the contrary, Sadness is produced because something bad has happened, so its wellbeing decreases. Fear appears when the possibility of something bad is about to happen. In this case, we expect that the wellbeing drops off. Finally, Anger is produced when a decrement of the wellbeing of the robot happened due to another-initiated act.

This paper presents a control architecture for an autonomous agent based on motivations. The agent uses reinforcement learning algorithms to learn its policy while interacts with the world. The reward for these learning algorithms will be the variation of the wellbeing of the agent (happiness/sadness) due

to the previous selected behaviour, calculated at each step of the process. This wellbeing is a function of the internal needs of the agent (drives). This idea of using the wellbeing of the agent as the reinforcement in the learning process for behaviour selection has been also used by Gadanho in the ALEC architecture, obtaining quite good results (Gadanho, 2003).

The remainder of the paper is organized as follows. Section 2 introduces the use of emotions in robots. Section 3 and 4 describe the proposed control architecture and the reinforcement learning algorithms respectively. Section 5 introduces the emotion fear and section 6 describes the experimental setting. Finally, conclusions and future works are summarized in section 7.

2 Emotions in robotic

One of the main objectives in robotics and artificial intelligence research is to imitate the human mind and behaviour. For this purpose the studies of psychologists on the working mind and the factors involved in the decision making are used. In fact, it has been proved that two highly cognitive actions are dependant not only on rules and laws, but on emotions: Decision making and perception (Picard, 1998). In fact, some authors affirm that emotions are generated through cognitive processes. Therefore emotions depend on ones interpretation, i.e. the same situation can produce different emotions on each agent, such as in a football match (Ortony, 1988). Moreover, emotions can be considered as part of a provision for ensuring and satisfaction of the system's major goals (Frijda, 1987).

Emotions play a very important role in human behaviour, communication and social interaction. Emotions also influence cognitive processes, particularly problem solving and decision making (Damasio, 1994). In recent years, emotion has increasingly been used in interface and robot design, primarily in recognition that people tend to treat computers as they treat other people.

There are several theories about emotions (Frijda 1987; Ortony, 1988; Sloman, 2003; Rolls, 2003), but the results of Damasio (1994) can be considered the basis, for many A.I. researchers, to justify the use of emotions in robotics and their computation.
Rosalind Picard in her book Affective Computing (1998), writes a complete dissertation about this subject based on several psychologists, including Damasio. Picard (1998) proposed a design criterion in order to create a computer that could express emotions. Moreover, she established that a computer has emotions if it has certain components that are present on the emotional systems of healthy people. Picard (2003) expounded four motives for giving certain emotional abilities to machines: The first goal is to build robots and synthetic characters that can emulate living humans and animals, such as a humanoid robot. The second is to make machines that are intelligent. A third objective is to try to understand human emotions by modelling them. Although these three goals are important, the main one is to make machines less frustrating to interact with, i.e. to facilitate the human-machine interface.

Cañamero (2003) considers that emotions, or at least a sub-group of them, are one of the mechanisms founded in biological agents to confront their environment. This creates ease of autonomy and adaptation. For this reason she considers that it could be useful to exploit this role of emotions to design mechanisms for an autonomous robot. Emotions are used as mechanisms that allow the agent (robot) to:

1. Have fast reactions.
2. Contribute to resolve the selection among multiple objectives.
3. Signal important events to others.

Bellman (2003) agrees, to some degree, with Cañamero and her reasons for considering emotions in robotics. The author states that emotions allow animals with emotions to survive better than the others without emotions. Therefore, we can presume that some type of analogy to emotional abilities is required within robots, if we want an intelligent and independent behaviour within a real environment.

Changing subject, Picard (2003) gives an advice about the implementation in machines of functions implemented by the human emotional system. Computers do not have emotions as human beings in any natural experimentation sense. Science methodology is to try to reduce complex phenomena, such as emotions, to a functional requirements list. The challenge of many computing science researchers is to try to duplicate these in computers at different levels depending on the motives of the investigation. But we must be careful when presenting this challenge to the general public, who may perceive that emotions are the frontier that separates man and machine

3 Control Architecture

An independent system should not have to wait for someone to maintain, succour, and help it (Frijda and Swagerman,1987). Therefore, an autonomous agent should be capable of determining its goals, and it must be capable of selecting the most suitable behaviour in order to reach its goals. Similarly to other authors (Avila-Garcia and Cañamero, 2004), (Breazeal, 2002), (Gadanho, 2003), (Velasquez, 1998), our agent's autonomy relies on a motiva-

tional model. Figure 1 shows this proposed control architecture for behaviour selection.

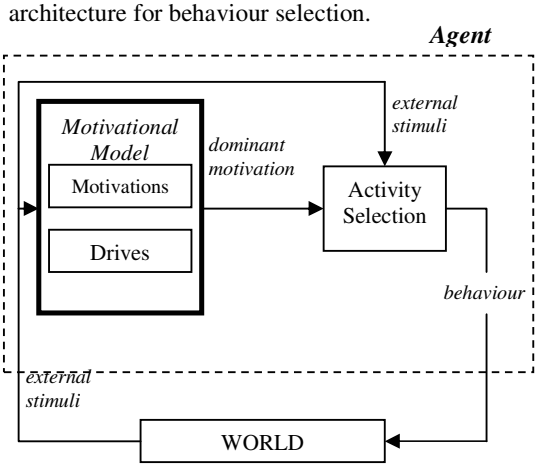

Figure 1: Control architecture for autonomous agents

3.1 Motivational Model

Motivations can be seen as homeostatic processes, which maintain a controlled physiological variable within a certain range. Homeostasis means maintaining a stable internal state (Berridge, 2004). This internal state can be parameterized by several variables, which must be around an ideal level. When the value of these variables differs from the ideal one, an error signal occurs: the drive. These drives constitute urges to action based on bodily needs related to self-sufficiency and survival. External stimuli, both innate and learned, are also able to motivate and drive behaviour (Cañamero, 1997).

In order to model motivation, the hydraulic model of motivation described by Lorentz and Leyhausen in (Lorentz and Leyhausen, 1973) has been used as an inspiration. This model is essentially a metaphor that suggests that motivational drive grows internally and operates a bit like pressure from a fluid reservoir that grows until it bursts through an outlet. Motivational stimuli from the external world act to open an outflow valve, releasing drive to be expressed in behaviour. In this model, internal drive strength interacts with external stimulus strength. If drive is low, then, a strong stimulus is needed to trigger motivated behaviour. If the drive is high, then, a mild stimulus is sufficient (Berridge, 2004). Following this idea, the intensity of motivations (M_i) is a combination of the intensity of the related drive (D_i) and the related external stimuli (w_i), as it is expressed in the following equation:

$$M_i = D_i + w_i \tag{1}$$

The ideal value for all the drives is 0. The external stimuli are the different objects that the player can find in the virtual world during the game. If the stimulus is present the value of w_i is 1, otherwise is 0.

According to (1), the intensity of a motivation is high due to two reasons: 1) the correspondent drive is high or 2) The correct stimulus is present. The dominant motivation is the one with the highest intensity.

This model can explain the fact that due to the availability of food in front of us, we sometimes eat although we are not hungry. We have also introduced activation levels (L_d) for motivations such that:

$$\begin{aligned} &\textit{if } D_i \le L_d \textit{ then } M_i = 0 \\ &\textit{if } D_i > L_d \textit{ then } (1) \text{ is applied} \end{aligned} \tag{2}$$

Therefore the possibility of no dominant motivation exists.

3.2 Wellbeing

As shown in (3), the agent's wellbeing is a function of the values of the drives (D_i) and some "personality" factors (α_i).

$$Wb = Wb_{ideal} - \sum_i \alpha_i \cdot D_i \tag{3}$$

Wb_{ideal} is the ideal value of the wellbeing of the agent, which is set to 100. The personality factors weight the importance of the values of the drives on the wellbeing of the agent. The value of the wellbeing and its variation (ΔWb) are calculated at each step. The variation of the wellbeing is calculated as the current value of the wellbeing minus the wellbeing value in the previous step.

3.3 Behaviour Selection

The action selection process consists in making decisions as to what behaviours to execute in order to satisfy internal goals and guarantee survival in a given environment and situation. For other authors (Avila and Cañamero, 2002), (Avila and Cañamero, 2004), (Cañamero, 1997) this implies that the agent can choose among some behaviors related to the dominant motivation. Therefore for each motivation there is a set of behaviours oriented to fulfill the motivational goal.

It is important to note that finally, the agent will learn that when the dominant motivation is Eat, it must select among the behaviours related to the object food, instead of those associated to water or medicine. The novelty of our approach is that these

behaviours were not linked a priori with the correspondent motivations.

3.4 Happiness and Sadness

Considering the definitions of the emotions given in the introduction section:

$$\begin{aligned} &If\ \Delta Wb > L_h \Rightarrow Happiness \\ &If\ \Delta Wb < L_s \Rightarrow Sadness \end{aligned} \quad (4)$$

Where $L_h>0$ and $L_s<0$ are the minimum variations of the wellbeing of the agent that produce Happiness or Sadness respectively. Therefore these two emotions are used by the agent as the reward for the reinforcement learning algorithms.

In this architecture the agent learns, using different reinforcement learning algorithms, the best behaviour at each step using happiness/sadness as the positive/negative reward. Therefore, in this architecture behaviours are not selected to satisfy the goals determined by the dominant motivation but to optimize the wellbeing of the agent. This implies that the final goal of the agent is to maximize Happiness.

4 Reinforcement Learning

Reinforcement learning (RL) is about learning from interaction how to behave in order to achieve a goal. The agent's objective is to maximize the amount of reward it receives over time (Sutton and Barto, 1998). Q-learning is a value learning version of RL that learns utility values (Q-values) of state and action pairs Q(s,a). It provides a simple way for agents to learn how to act optimally in controlled Markovian domains (Yang and Gu, 2004). The theory of Markov Decision Processes (MDP's), assumes that the agent's environment is stationary and as such contains no other adaptive agents (Littman, 1994). Therefore, while the agent is not interacting with the other agent, we will consider our virtual world as a MDP environment.

On the other hand, if the agent is interacting with other player, the rewards the agent receives depend not only on their own actions but also on the action of the other agent. Therefore, the individual Q-learning methods are unable to model the dynamics of simultaneous learners in the shared environment. Currently multiagent learning has focused on the theoretic framework of Stochastic Games (SGs) or Markov Games (MGs). SGs appear to be a natural and powerful extension of MDPs to multiagent domains (Yang and Gu, 2004).

Taking into account these considerations, in the proposed architecture the agent will use the standard Q-learning algorithm as the RL algorithm when the agent is not interacting with the other player. Obviously, in the case of "social" interaction, the agent must use a multiagent RL algorithm. The following subsections explain in more details these two scenarios.

In our system, the state of the agent is the aggregation of his inner state S_{inner} and the states S_{obj} related to each of the objects, including external agents, which can interact with him.

$$S = S_{inner} \times S_{obj_1} \times S_{obj_2} \ldots \quad (5)$$

For the RL algorithms the states related to the objects are considered as independent. This means that the state of the agent in relation with each object is $s \in S_{inner} \times S_{obj_i}$

4.1 Q-learning Algorithm

As mentioned previously, in MDP environments the agent will use the standard Q-Learning as a learning algorithm. As described in (Gadanho, 2002), through this algorithm the agent learns iteratively by trial and error the expected discounted cumulative reinforcement that it will receive after executing an action a in response to a world state s, the Q-values for each object is:

$$Q^{obj_i}(s,a) = (1-\alpha)\cdot Q^{obj_i}(s,a) + \alpha\cdot\left(r + \gamma \max_{a \in A_{obj_i}} \left(Q^{obj_i}(s',a)\right)\right) \quad (6)$$

where A_{obj_i} is the set of actions related to the object i, s' is the new state, r is the reinforcement; γ is the discount factor and α is the learning rate parameter.

The optimal policy, chooses the action that maximizes $Q^{obj_i}(s,a)$ this means

$$a^* = \arg\max_a Q^{obj_i}(s,a) \quad (7)$$

The proposed architecture differs from others in that we do not consider only the behaviours that help to satisfy the drive related with the dominant motivation but the agent must consider all the behaviours that can be performed at each step, depending on his states.

4.2 Multiagent reinforcement learning

In multiagent systems, other adapting agents make the environment no longer stationary so the Markov property is not applicable. In the learning framework of SGs, learning agents attempt to maximize their expected sum of discounted rewards. Unlike single-agent system, in multiagent systems the joint actions determine the next state and rewards of each agent. In (Littman, 1994) it is proposed a Minimax-

Q learning algorithm for zero-sum games in which the player always tries to maximize its expected value in the face of the worst-possible action choice of the opponent. The player's interests in the game are opposite. Later, Littman (Littman, 2001) proposed the Friend or Foe Q-learning algorithm, for the RL in general-sum SGs. The main idea is that each agent is identified in advance as being either "friend" or "foe". The Friend class consists of SGs in which the Q-values of the players define a game which has a coordination equilibrium. The Foe class is the one in which the Q-values define a game with an adversarial equilibrium. The Friend-Q updates similarly to regular Q-learning, and Foe-Q updates as does minimax-Q (Shoham et al, 2003).

All these algorithms extend the normal Q-function of state-action pairs $Q^{obj_i}(s,a)$ to a function of states and joint actions of all agents. Taking into account this fact and that each agent can select among n actions while they are interacting, the Q-values to be calculated are $Q^{obj_i}(s,a_1,a_2)$ where a_1 and a_2 belong to the set of n actions of each agent.

5 Fear

Fear is produced when the agent knows that something bad may happen. This means that the wellbeing of the agent might decrease. To cope with fear the action that produces the negative effect is going to be considered. We will distinguish between actions executed by the agent and exogenous actions carried out by other elements of the environment such as other agents.

5.1 To be afraid of executing risky actions

Q-learning algorithm evaluates every action carried out in a state, using the expected average value. However, since the system is non deterministic, the result of a certain action may have different values. The worst result experimented by the agent for each pair action-state is stored in a variable called $Q^{obj_i}_{worst}(s,a)$, which is updated after the execution of the action.

$$Q^{obj_i}_{worst}(s,a) = \min(\, Q^{obj_i}_{worst}(\, s,a),\; r + \gamma \max_{a \in A_{obj_i}} (Q^{obj_i}(s',a))) \quad (8)$$

where A_{obj_i} is the set of actions, *s'* is the new state, *r* is the reinforcement and γ is the discount factor.

The effect of being afraid can be considered by choosing the action that maximizes $Q^{obj_i}_{fear}$ instead of choosing the one that maximizes Q^{obj_i},

$$Q^{obj_i}_{fear}(s,a) = \beta Q^{obj_i}(\, s,a) + (1-\beta) Q^{obj_i}_{worst}(\, s,a) \quad (9)$$

Using this approach the expected result of each action is considered as well as the less favourable one. The parameter β, being $0 \leq \beta \leq 1$, measures the daring degree of the agent, and its value will depend on the personality of the agent. If the agent is fearless, β will be near 1; while in a fearful agent, who tries to minimize the risk, β will be near 0. If $\beta = 1$ the agent is using the optimal policy.

This means that the "fearful" policy chooses the action:

$$a^f = \arg\max_a Q^{obj_i}_{fear}(\, s,a) \quad (10)$$

For example, when an agent has to pass over a deep hole, he can choose between jumping over it and going around it. Jumping is easier, faster and usually safe, but very occasionally he can fail and die. On the other hand, if the agent goes around the hole he will take a lot of time and get tired but it is safer. Translating this example to our point of view, the Q-value related with jumping will be greater than the one related to going around. Using the standard Q-learning algorithm, the agent would always jump over the hole. Using the fearful policy, considering the worst thing that could happen to the agent jumping or going around, he would choose going around since it is safer than jumping.

5.2 To be afraid of malicious exogenous actions

When the agent may suffer some negative effects in a state as a consequence of exogenous events, feels fear. "Fear" is expressed as a drive D_{fear}.

Traditionally, Q-learning has been applied on Markov decision processes (MDP), which are discrete time systems. Some authors have extended the use of this algorithm to continuous time systems by considering them as semi Markov decision processes. In both cases it is commonly assumed that there are no exogenous events. In order to introduce the effects of exogenous events in continuous systems we consider the system as a discrete time system with constant period. In the limit, if the period is very small the system will tend to be a continuous time system. Moreover, we will also consider that the exogenous events can be associated to other agents or elements of the environment. These exogenous events are synchronized with the actions executed by the agent. Among these action we will include the action of "doing nothing". In this case the treatment for multiagent systems mentioned before will be applied.

The exogenous events executed by an external object or other agent can occur simultaneously to any of the actions of the agent. Therefore the negative effects of these exogenous events will be reflected in all the actions of the agent. In order to separate the effects of the actions of the agent and the effects of the exogenous events, we will focus on the study of the agent when he is "doing nothing". In that case, we suppose that all the changes suffered by the agent are a consequence of external elements.

It will be considered that a state is a "scary" state when:

$$Q_{worst}^{obj_i}(s, Nothing) < L_{fear} \quad (11)$$

being L_{fear} the minimum acceptable value of the worst result that can be expected by the agent when it is doing nothing. In this case the value of the fear drive D_{fear} will be incremented.
When

$$Q_{worst}^{obj_i}(s, Nothing) > L_{safe} \quad (12)$$

it is considered that the agent is in a "safe" state and the value of the fear drive D_{fear} will be decreased.

The fear drive is equally treated as the rest of drives, and its related motivation could be the dominant one. In this case, the agent will learn by itself what to do when it is afraid.

6 Experimental Test Bed

The proposed architecture is intended to be used in a social personal robot developed by our lab and named "Maggie" (see Fig2) (Salichs et al, 2006). As a first stage of this project and due to the obvious physical difficulties of making experiments on a real robot and on a real environment, we decided to implement our architecture on virtual players, who "live" in a virtual world, a text-based multi user role game. This game gave us the possibly of creating different 2-D environments to play in, as well as a graphic interface.

Table 1 shows our agent's motivations, drives and external stimuli that the agent can find in the virtual world.

These drives have been selected tacking into account the role of the agent in the virtual world used to implement our architecture. Since our final goal is to construct an autonomous social robot, it must show social behaviours. Therefore, as it is shown, social motivations are included as robot's needs.

Table 1: Motivations, drives and related stimuli

Drive/Motivation	**External Stimuli**
Energy	Food
Thirst	Water
Health	Medicine
Sociability	Other player
Fear	

At each simulation step some of these drives, such as Energy, Thirst, Health and Sociability are incremented by a certain amount. The value of the drive Fear, as it was previously explained, increase or decrease depending on if the agent is in a "scary" state or not.

Following (3) the wellbeing of the agent is defined by:

$$Wb = Wb_{ideal} - (\alpha_1 D_{energy} + \alpha_2 D_{thirst} + \alpha_3 D_{health} + \alpha_4 D_{social} + \alpha_5 D_{fear}) \quad (13)$$

In our test bed the inner state is then:

$$S_{inner} = \{Hungry, Thirsty, Ill, Bored, Scary, OK\} \quad (14)$$

This internal state is obviously related with the dominant motivation. Therefore when the dominant motivation is for example "Eat" then the agent is "Hungry" and so on.

In relation with static objects the agent can be in the following states:

$$S_{obj} = Have\_it \times Near\_of \times Know\_where \quad (15)$$

where,

$$Have\_it = \{yes, no\} \quad (16)$$

$$Near\_of = \{yes, no\} \quad (17)$$

$$Know\_where = \{yes, no\} \quad (18)$$

In relation with other player:

$$S_{obj} = Near\_of \quad (19)$$

where,

$$Near\_of = \{yes, no\} \quad (20)$$

And the set of actions that can be executed in every state is the following:

$$A_{food} = \{Eat, Get, Go\_to, Explore\} \quad (21)$$

$$A_{water} = \{Drink, Get, Go\_to, Explore\} \quad (22)$$

$$A_{medicine} = \{Take, Get, Go\_to, Explore\} \quad (23)$$

$$A_{playmate} = \begin{cases} Explore \\ Steal\ \ food / water / medicine \\ Give\ \ food / water / medicine \\ Chat \end{cases} \quad (25)$$

Among the previously mentioned behaviours there are some of them that reduce or increase some drives, and therefore will produce a variation in the emotional state of the agent:

- Eat food: reduces to zero the Energy drive. (happiness when hungry)
- Drink water: reduces to zero the Thirst drive. (happiness when thirsty)
- Take medicine: reduces to zero the Health drive. (happiness when sick)
- Chat: reduces to zero the Social drive. (happiness when the social drive is high)
- To be taken something by other player: increases by a certain amount the Social drive. (sadness)
- To be given something from other player: reduces by a certain amount the Social drive. (happiness when the social drive is high)

Fig. 2. "Maggie" The Social Robot of the Robotic Lab.

The conducted experiments show the usefulness of the proposed architecture in facilitating the development of social autonomous agents able to learn from the experience the right behaviours to execute depending on the world state.

7 Conclusion and Future work

In this paper different reinforcement learning algorithms have been discussed and implemented for the behaviour-selection learning of non-interacting and social autonomous agents. These agents are controlled by an emotion-based architecture, which performs homeostatic control of the internal state of the agent through an embedded motivational model. This architecture has been designed for autonomous and social robots.

The agent is intrinsically motivated and his goal is his own wellbeing. The learning algorithms use happiness/sadness of the agent as positive/negative reinforcement signals. Fear is used to prevent the agent choosing dangerous actions or being in dangerous states where non-controlled exogenous events, produced by external objects or other agents, could danger him.

In the future work, it is expected that the agent learns not only the right policy but also to identify its opponent. So far, the agent treats all its opponents as if they were all the same, and this is not true. In future scenarios, the agent will be able to behave different with the "good" opponent than with the one that tries to steal its objects every time that interacts with it.

Another emotion is going to be implemented: Anger. Anger will be produced when sadness arises due to the interaction with another agent

Acknowledgements

The authors gratefully acknowledge the funds provided by the Spanish Government through the projects named "Personal Robotic Assistant" (PRA) and "Peer to Peer Robot-Human Interaction" (R2H), of MEC (Ministry of Science and Education).

References

Avila-Garcia, O. and Cañamero, L. A Comparison of Behavior Selection Architectures Using Viability Indicators. In *Proc. InternationalWorkshop Biologically-Inspired Robotics: The Legacy of W. Grey Walter*(WGW'02). 2002.

Avila-Garcia, O. and Cañamero, L. Using Hormonal Feedback to Modulate Action Selection in a Competitive Scenario. *In Proc. 8th Intl.Conference on Simulation of Adaptive Behavior (SAB'04)*. 2004

Bellman , Kirstie L.. Emotions: Meaningful mappings between the individual and its world. In: *Emotions in Humans and Artifacts*. (Robert Trappl, Paolo Petta and Sabine Payr), pp 149-188. The MIT Press. Cambridge, Massachusetts. 2003

Berridge , Kent C. Motivation concepts in behavioural neuroscience. *Physiology & Behaviour* 81, 179-209,2004,.

Breazeal C. *Designing Sociable Robots*. The MIT Press. 2002

Cañamero, D. Modeling Motivations and Emotions as a Basis for Intelligent Behavior. *In W. Lewis Johnson, ed., Proceedings of the First International Symposium on Autonomous Agents (Agents'97)*, 148-155. New York, NY: The ACM Press. 1997.

Cañamero, D. Designing emotions for activity selection in autonomous agents. In: *Emotions in Humans and Artifacts.* (Robert Trappl, Paolo Petta and Sabine Payr), pp 115- 148. The MIT Press. Cambridge, Massachusetts. 2003

Damasio, Antonio. *Descartes' Error – Emotion, reason and human brain*. Picador, London. 1994

Fong. T., Nourbakhsh. I., Dautenhahn K. *A survey of socially interactive robots: Concepts, design, and applications.* Technical Report CMU-RI-TR-02-29. 2002

Frijda, N. and Swagerman, J. Can computers feel? Theory and design of an emotional model. *Cognition and Emotion.* 1 (3). pp 235-357. 1987

Fujita, Masahiro AIBO: Toward the Era of Digital Creatures. *The International Journal of Robotics Research*. Vol 20, Nº 10, pp 781-794. October 2001

Gadanho, Sandra Clara. Emotional and Cognitive Adaptation in Real Environments. *In: Symposium ACE'2002 of the 16th European Meeting on Cybernetics and Systems Research*, Vienna, Austria. 2002

Gadanho, Sandra Clara. Learning behavior-selection by emotions and cognition in a multi-goal robot task. *The Journal of Machine Learning Research.* Volume 4 Pages: 385 – 412. MIT Press Cambridge, MA, USA. 2003

Littman, M. L. Markov games as a framework for multiagent learning. *In Proceedings of the Eleventh International Conference on Machine Learning*, San Francisco, California, pp. 157--163. 1994

Littman, M. L. Friend-or-foe Q-learning in general-sum games. *In Proceedings of the Eighteenth International Conference on Machine Learning*, pages 322--328, Williams College, June 2001.

Lorentz K, Leyhausen P. *Motivation of human and animal behaviour; an ethological view*. New York: Van Nostrand-Reinhold; xix, 423 pp. 1973

Malfaz, M. and Salichs, M.A.. A new architecture for autonomous robots based on emotions. *Fifth IFAC Symposium on Intelligent Autonomous Vehicles*. Lisbon. Portugal. Jul, 2004.

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

Picard, Rosalind W. *Affective computing.* Ed. Ariel S.A. 1998

Picard, Rosalind W. What does it mean for a computer to have emotions?. In. *Emotions in Humans and Artifacts.* (Robert Trappl, Paolo Petta and Sabine Payr), pp 213-235. The MIT Press. Cambridge, Massachusetts. 2003

Rolls, Edmund, T. A Theory of emotion, its functions, and its adaptive value. In: *Emotions in Humans and Artifacts.* (Robert Trappl, Paolo Petta and Sabine Payr), pp 11-35. The MIT Press. Cambridge, Massachusetts. 2003

Salichs, M. et al. Maggie: A Robotic Platform for Human-Robot Social Interaction. In *2006 IEEE International Conferences on Cybernetics & Intelligent Systems (CIS) and Robotics, Automation & Mechatronics.* Bangkok, Thailand. 2006

Shibata T., Tashima T.,Arao M., Tanie K. Interpretation in Phisical Interaction between human and artificial emotional creature. *Proceedings of the 1999 IEEE. International Workshop on Robot and Human Interaction*. Pisa, Italy – September 1999

Shoham, Y., Powers, R. and Grenager, T. *Multi-agent reinforcement learning: a critical survey.* Technical report, Computer Science Department, Stanford University, Stanford. 2003.

Sloman, Aaron. How many separately evolved emotional beasties live within us. In: *Emotions in Humans and Artifacts*. (Robert Trappl, Paolo Petta and Sabine Payr), pp 35-115. The MIT Press. Cambridge, Massachusetts. 2003

Sutton, Richard S. and Barto, Andrew G. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, A Bradford Book.1998

Velásquez, J. When Robots Weep: Emotional Memories and Decision Making. *In: Proceedings of AAAI-98*. 1998

Yang E. and Gu D. *Multiagent Reinforcement Learning for Multi-Robot Systems: A Survey.* Technical Report CSM-404.University of Essex. (2004)

Towards a Computational Account of Reflexive Consciousness

Murray Shanahan

Department of Computing, Imperial College London, 180 Queen's Gate, London SW7 2AZ, UK.
m.shanahan@imperial.ac.uk

Abstract

This paper offers a preliminary sketch for an account of reflexive consciousness based on an implemented architecture that combines a global workspace architecture with an internally closed sensorimotor loop. The proposed account extends the theoretical framework of the already implemented architecture with to two concepts that structure the flow of consciously processed information. First, contextual switches divide the unfolding contents of consciousness into a set of nested episodes, wherein one conscious episode can "refer to" another. Second, the imposition of a focus / fringe structure enables consciousness to encompass material that is merely available to it but not actually present. This combination of reflexivity and fringe may underpin our awareness of our own existence as conscious beings.

1 Introduction

Cognitive theories of consciousness, as the name suggests, posit an intimate link between cognition and consciousness. For example, according to *global workspace theory* (Baars, 1988; 1997; 2002), non-conscious information processing in the human brain is carried out locally within specialist brain processes, while the hallmark of consciously processed information is that it is broadcast (via a "global workspace") and made available to the entire set of these specialists. The upshot is that consciously processed information is cognitively efficacious in ways that non-consciously processed information is not. Specifically, the procession of broadcast global workspace states resembles a serial thread of computation, yet it integrates the results of massively parallel computation, sifting out relevant contributions from the irrelevant (Shanahan & Baars, 2005).

However, one feature of conscious human thought not accounted for by global workspace theory in its basic guise is *reflexivity*, that is to say the capacity for a conscious thought to refer to itself or to other conscious states. (By contrast, so-called higher-order thought (HOT) theories of consciousness take reflexivity as their primary datum (Rosenthal, 1986).) If consciously processed information is, as global workspace theory maintains, cognitively efficacious, then reflexively conscious information processing is even more so – since it enables the thinking subject to reflect on his or her own mental operations, to critique them and improve on them, and to respond to the ongoing situation in ways that depend on a degree of self-knowledge. So the question arises: Can global workspace theory be extended to account for reflexive consciousness?

This question has phenomenological as well as cognitive implications. For if we accept the argument of Shanahan (2005), the very idea of a conscious subject – something it is like something to be, in Nagel's well-known terminology – can be objectively accounted for in terms of a suitably *embodied* instantiation of the global workspace architecture, wherein all the specialist processes are indexically directed towards maintaining the wellbeing and fulfilling the purpose (or "mission") of a single, spatially unified body. By extending global workspace theory to reflexive consciousness, we can bolster this line of argument by showing that a similar treatment is available for a vital aspect of human phenomenology, namely our ability to become conscious of our own existence *as* conscious subjects.

2 Internal Simulation with a Global Workspace

Figure 1. illustrates the operation of the global workspace architecture, which comprises a set of specialist brain processes plus a global workspace. Information processing within the architecture consists of periods of *competition* interleaved with periods of *broadcast*. On the left of the figure, we see the set of specialist processes competing to gain access to the global workspace. Gaining access entails that the winning process (or coalition of processes) gets to broadcast its message, via the global

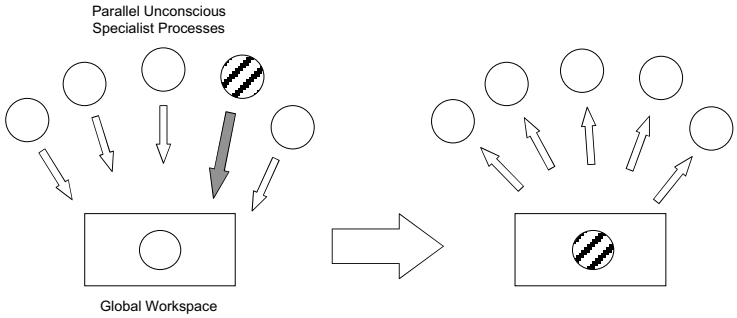

Fig. 1: The Global Workspace Architecture.

workspace, back to the entire set of specialists, as seen on the right of the figure. The global workspace itself is, in essence, nothing more than the infrastructure of a communications network that permits signals generated within localised neuronal populations to influence remote, widespread brain regions. According to global workspace theory, the mammalian brain instantiates such an architecture, and this allows us draw a empirically falsifiable distinction between consciously and non-consciously processed information. Information processing that is confined to local specialists is necessarily non-conscious, and only broadcast information can be consciously processed.

Although global workspace architecture permits this fundamental distinction to be drawn in a theoretically respectable manner, it still leaves open the question of the content of consciously processed information. But by augmenting the basic global workspace architecture with an internally closed sensorimotor loop (Fig. 2), it is possible to reconcile it with another idea current within the scientific study of consciousness, namely the *simulation hypothesis*, according to which thought is internally simulated interaction with the environment (Cotterill, 1998; Hesslow, 2002; Shanahan, 2006). If the sophisticated mental life of a human being results from the interplay of external stimulation with internally generated activity such as inner speech and mental imagery, then something like the internally closed sensorimotor loop posited by the simulation hypothesis is required to account for it. Moreover, by facilitating the *rehearsal* of trajectories through sensorimotor space, the internal sensorimotor loop helps the individual to anticipate the consequences of their actions and to plan ahead, and thereby fulfils a fundamental cognitive role.

In (Shanahan, 2006), a implemented system is described that reconciles global workspace theory with the simulation hypothesis. The system controls a simple two-wheeled robot with a camera, and enables it to select an action based not only on a set of reactive responses, but also taking into consideration the result of simulating the expected outcomes of its actions using an internal sensorimotor loop, as depicted in Figure 3. Moreover, a global workspace is incorporated into the loop. The procession of states exhibitted by the global workspace, which simulates a possible trajectory through the robot's sensorimotor space, is the outcome of both competition and broadcast : the i^{th} state being broadcast to multiple neuronal populations which then compete to determine the $i+1^{th}$ state. Further details of the system are beyond the scope of this article, and can be found in (Shanahan, 2006).

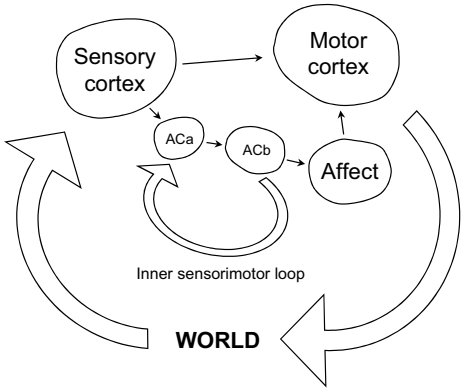

Fig. 2: External and Internal Sensorimotor Loops

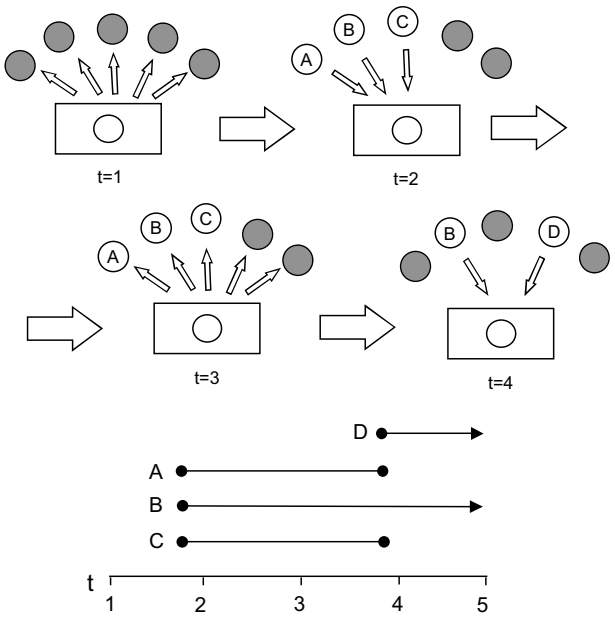

Fig. 3: The Temporal Structure of Consciously Processed Information

3 Context and Temporal Structure

According to the account of reflexive consciousness proposed by this paper, the flow of information through the global workspace is divided into distinct, but possibly nested or overlapping, episodes at various timescales. Beginnings and ends of conscious episodes are triggered by events (contextual switches) – such as entering and leaving a room, or meeting and parting from a friend – which wake up or send to sleep relevant specialist processes, whose job it is to manage the individual's response to situations of that particular type. Figure 3 illustrates the idea. On the left of the figure are snapshots of the global workspace (GW) at four time points. At t=1, all five of the processes depicted are dormant, although they are still receiving information broadcast from GW. This information indicates the occurrence of a distinctive event – a *contextual switch* – and by t=2 this has caused three of the processes (A, B, and C) to become active and begin competing for access to GW. There follows a further period of broadcast (t=3), indicating a new contextual switch. By t=4, this has caused processes A and C to go back to sleep, but has woken up process D.

As Figure 3 shows, competition for access to GW is restricted to the currently active or "awake" set of specialist processes, and the set of active processes can be thought of as reflecting the current *context* (Fig. 3, top), a conception which is broadly in line with the notion of context prominently deployed by Baars (1988) in his original presentation of global workspace theory. Each distinct conscious episode, bracketed by a pair of contextual switches, falls under the jurisdiction of a particular process, a process that should be relevant in the current context. Intuitively, temporal context is a richly structured, hierarchical concept. The context of a lunchtime falls within the larger context of a day, while the context of a conversation can overlap the context of a lunchtime. Similarly, conscious episodes, which are associated with temporal contexts, can be nested or overlapping. However, it should be noted that diagrams such as Figure 3 (bottom) only show the set of processes that have the *potential* to contribute to the unfolding content of the global workspace at any given time point. For example, although process B is *active* at time t=3 in Figure 3, this does not entail that it has won (full) *access* to GW at time t=3. This means that (focal) consciousness typically does not contribute to a conscious episode for its entire duration, but only at those times when the corresponding process gains access to GW. On the other hand, as we'll see in the next section, any active process competing for access can contribute to *fringe* consciousness.

Allowing specialist processes to wake up and go to sleep in response to contextual cues gives them a simple form of internal state (on or off), and therefore allows them to respond to information in a way that is sensitive to past events. But from the standpoint of the present paper, the most important consequence of this demarcation of conscious episodes is

that it allows one such episode to "refer to" another. This could occur either when the referring episode of conscious thought falls entirely within the episode it is referring to (Fig. 4, left), or when the referred-to episode of conscious thought and the referring episode of conscious thought both occur within a third, enclosing conscious episode and the former occurs before the latter (Fig. 4, right). In either case, the referred-to episode might be the an ongoing experience, the recollection of an experience from the distant past (long-term memory) or the recent past (working memory), or part of an ongoing rational or creative process involving inner rehearsal. A typical referring (reflexively conscious) episode might offer some judgement on the (non-reflexively conscious) episode it is referring to, such as "that was unpleasant" or (for a reasoning process) "that hasn't got me any further".

4 Focus and Fringe

The above characterisation of a reflexively conscious thought as a conscious episode that "refers to" another conscious episode is all very well. But it leaves open many questions, including that of the mechanism by which this reference is achieved. So to flesh out our account of reflexivity, something further is required. According to the present treatment, in addition to the temporal structure described above, the flow of consciously processed information has a focus / fringe structure (Mangan, 1993; 2001). The fringe contains hints of material that has the potential to be brought into focal consciousness if required. As Mangan (2001) puts it, "The fringe creates a non-sensory feeling of imminence which implies the existence of far more than consciousness actually presents at any given moment. ... This is the fundamental trick that lets consciousness finesse its severely limited capacity ...".

The contention of this paper is that this is indeed a "fundamental trick", a means to enhance the cognitive efficacy of conscious information processing in many ways. Of especial interest here is the fact that, at any given time, while focal consciousness is contributing to one conscious episode, broadcasting information supplied by the corresponding active process, the fringe can simultaneously retain the trace of *another* co-occurring conscious episode, governed by a different active process. To see this, consider Figure 4 (right). Suppose that at time t=3 active process Z has won access to GW, and is therefore supplying the current content of focal consciousness. At the same time, although process Y is not enjoying (full) access to GW, it is still active, and can therefore influence fringe consciousness.

We have the outline, here, of mechanism by which one conscious episode can refer to another, wherein the referring episode is in focal consciousness while fringe consciousness retains a trace of the referred-to episode. But to see how this might be realised more concretely we need to zoom in and examine the evolving contents of GW at a finer timescale. In the computer model described in (Shanahan, 2006), GW was implemented as an attractor network. During execution, GW exhibited periods of stability (broadcast) during which it settled into an attractor, punctuated by periods of rapid change (competition) during which it got nudged out of a previously stable attractor and taken into a new one. During the periods of competition, it was sometimes observed that faint hints of competing attractors would become temporarily overlaid on GW's current attractor, each trying to take over.

This suggests the possibility that fringe consciousness might be realised as a rapid series of *faintly pulsing attractors*, each of which becomes transiently overlaid on the current attractor, but none of which yet has enough influence to dominate GW completely. (The dynamics here is reminiscent of Bressler & Kelso's (2001) notion of *metastability*.) Because these brief attractor pulses occur in GW, they are broadcast, and can therefore contribute to the flow of conscious information, as global workspace theory requires. Now we can appeal to *temporal synchrony*, as postulated by various authors as a solution to the binding problem (von der Malsburg, 1999), to realise reference between conscious episodes. The process currently supplying the content of focal consciousness – that is to say, the process associated with the referring episode – simply has to wait for the attractor corresponding to the process associated with the referred-to episode to pulse in

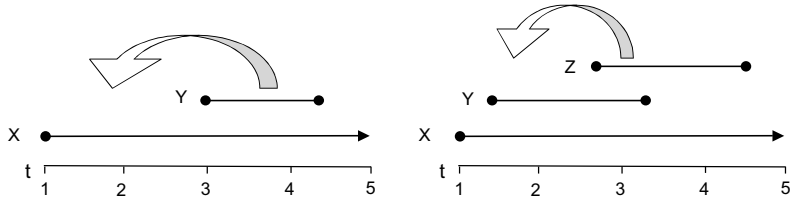

Fig. 4: Reflexively Conscious Episodes

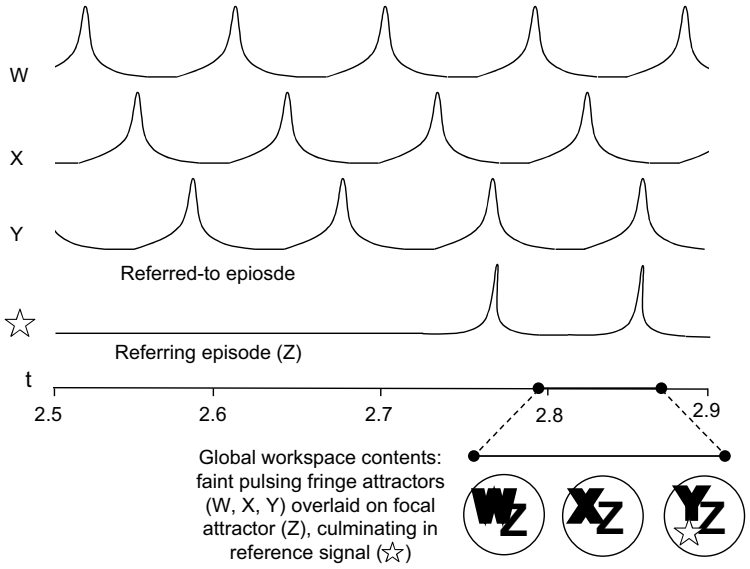

Fig. 5: Focus-Fringe Reference by Temporal Synchrony

GW, when it can, so to speak, signal "THAT ONE" to GW (Fig. 5). Since this signal will be broadcast at the same time as the attractor pulse of the referred-to episode, the required reference will be secured.

5 Fringe-Borne Self-Awareness

According to the simulation hypothesis, conscious thought is simulated interaction with the environment. This entails that insofar as a conscious experience relates to anything other than an immediately present stimulus, the information processing that underpins it, as well as implicating the broadcast mechanism of the global workspace, must recruit a higher-order, internally closed sensorimotor loop (Fig. 2). This is the case for both the recall of a past conscious episode and the conscious rehearsal (or imagination) of a trajectory through sensorimotor space, where the latter conception encompasses inner speech, mental imagery, and so on.

Now, the fundamental role posited for the fringe is to augment the flow of consciously processed information with an awareness of the many possible ways that the content of the GW could unfold from its present state, without having to supply detailed information about any one of those possibilities. For example, our awareness of the three-dimensionality of a solid object can be cashed out in terms of a fringe awareness of a host of sensorimotor possibilities, such as moving around to view the back of the object, or picking it up and rotating it to see a different facet.

In the context of an internal sensorimotor loop, the fringe carries an awareness of the tree of possiblilities for conscious recall or rehearsal that branches out from the GW's current state (Fig. 6). Now suppose that, using the mechanism outlined in the previous section, one (reflexively) conscious episode Z refers to another conscious episode Y with the thought "that didn't work because P" (in the case of recall) or "that wouldn't work because P" (in the case of rehearsal). Then, thanks to the broadcast of this message, the entire set of specialist, unconscious processes will be offered the challenge of finding a potential variation of Y in which P is not the case. The vast majority of these specialists will be irrelevant to Y. But any that are successful in finding a potentially useful variation will be able to promote, via the fringe, the possibility of rehearsing it properly. This shows how reflexive consciousness can marshall massively parallel resources to further increase the cognitive power of (non-reflexive) conscious information processing, which is itself more cognitively efficacious than non-conscious information processing.

To round off the account, let's develop further the parallel between fringe-borne spatial awareness (of solid objects, for example), and the fringe-borne awareness of the unfolding content of the global workspace itself. According to the present account, the conscious awareness of the three-dimensionality of nearby objects or of the space through which the body can move consists of hints in the fringe of a systematically organised set of possible trajectories through sensorimotor space. These hints are systematically organised in the sense that they conform to various constraints, which include the reversibility of certain actions (eg: moving forwards then backwards gets you back where you started) and the cyclic character of certain trajectories (eg: turning an object

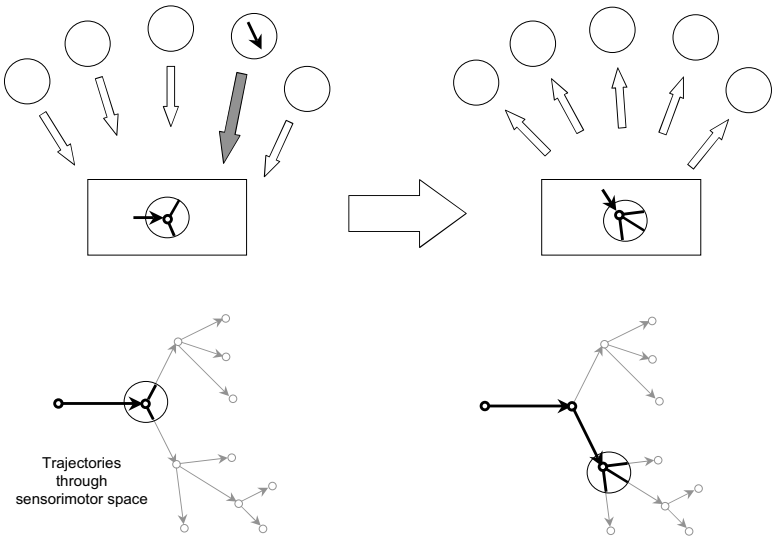

Fig. 6: Fringe-Borne Awareness of Possible Sensorimotor Trajectories

through 360° takes it back to its initial configuration).

In a similar vein, the fringe may sustain our awareness of the personhood of both ourselves and of others, hinting at material available for conscious rehearsal that pertains to our or their bodies, biographies, likes and dislikes, beliefs, desires, and intentions, skills and abilities, and so on. In the present context, the portion of this fringe-borne material of most interest relates to the way the content of the individual's consciousness unfolds. As the fringe-borne awareness of an object's solidity implies awareness of a systematic set of spatial constraints, so the fringe-borne awareness of personhood implies awareness of a systematic set of constraints on consciousness, such as its unity, its identity over time, and its indexical relationship to the body. Furthermore, in the same way that spatial constraints govern conscious thinking about solid objects, so these phenomenological constraints govern reflexively conscious thought. Insofar as we become consciously aware of ourselves as conscious beings, perhaps we do so thanks to our capacity to entertain reflexive thoughts combined with a fringe-borne awareness of the laws governing the way conscious thought unfolds.

References

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

Baars, B.J. (1997). *In the Theater of Consciousness: The Workspace of the Mind*. Oxford University Press.

Baars, B.J. (2002). The Conscious Access Hypothesis: Origins and Recent Evidence. *Trends in Cognitive Science* 6 (1), 47–52.

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

Cotterill, R. (1998). *Enchanted Looms: Conscious Networks in Brains and Computers*. Cambridge University Press.

Hesslow, G. (2002). Conscious Thought as Simulation of Behaviour and Perception. *Trends in Cognitive Science* 6 (6), 242–247.

Mangan, B. (1993). Taking Phenomenology Seriously: The "Fringe" and its Implications for Cognitive Research. *Consciousness and Cognition* 2 (2), 89–108.

Mangan, B. (2001). Sensation's Ghost: The Non-Sensory "Fringe" of Consciousness. *PSYCHE* 7 (18), http://psyche.cs.monash.edu.au/v7/psyche-7-18-mangan.html.

Rosenthal, D. (1986). Two Concepts of Consciousness. *Philosophical Studies* 49 (3), 329–359.

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

Shanahan, M.P. (2005). Global Access, Embodiment, and the Conscious Subject. *Journal of Consciousness Studies* 12 (12), 46–66.

Shanahan, M.P. (2006). A Cognitive Architecture that Combines Internal Simulation with a Global Workspace. *Consciousness and Cognition*, in press.

Von der Malsburg, C. (1999). The What and Why of Binding: A Modeler's Perspective. *Neuron* 25, 95–104.

How to experience the world: some not so simple ways

Aaron Sloman
School of Computer Science, University of Birmingham,
Edgbaston, Birmingham, B15 2TT, UK
A.Sloman@cs.bham.ac.uk
http://www.cs.bham.ac.uk/~axs/

Extended Abstract:

I believe the best way to extend our scientific understanding of consciousness is to stop using the noun and investigate all the many mental processes that can and do occur in humans and other animals and future robots in very great detail and explain how they are possible. Then everything of substance about consciousness will have been covered, and the vacuous, incoherent unanswered questions generated in philosophical discussions will remain unanswered as they should be, because they are unanswerable.

My talk is an illustration of a small part of this project, starting from a comment made by Wittgenstein when discussing the experience of ambiguous figures. He wrote:

> The substratum of this experience is the mastery of a technique.

I don't really know what he meant by that, but those words slightly modified thus:

> The substratum of an experience is mastery of a large collection of techniques available and ready to be deployed if required, possibly in new combinations.

could be used to express a theory I am trying to develop in the context of trying to understand how to give a robot human-like (to be more precise, child-like) capabilities in the context of perceiving and manipulating 3-D objects.

The idea is that an infant-toddler-child-youth (and future domestic robot) develops by constantly actively and creatively exploring many aspects of the environment and thereby learning a very large number (possibly many thousands, certainly many hundreds) of different facts about the environment including facts about different kinds of stuff things are made of, different kinds of surface fragments that can occur, different kinds of ways things can be combined or decomposed, different kinds of relationships that can occur between simple and complex objects, different ways collections of relations can change, different kinds of actions that can be produced, and of course different consequences of all the above.

These facts are not expressed as propositions using what we would call a human language, but they must be somehow represented internally in a usable form, and in particular, for creative experiments to be performed and novel problems to be solved by combining prior knowledge the information must be recombinable in novel ways for some uses.

So, a child or future intelligent domestic robot is constantly learning orthogonal, recombinable, competences. (Actually, not totally orthogonal since independent variation of phenomena is limited in many ways, that have to be learnt.) It seems that precocial species either cannot do this or do it to a much more limited extent: they start off with the vast majority of what they need to know about the world and how to act in it pre-programmed by evolution (contradicting familiar arguments about the requirements for 'symbol grounding'). Altricial species that develop very complex and diverse cognitive competences probably evolved these powerful information acquiring, restructuring, mechanisms because (a) genetic mechanisms lacked the space to encode them and (b) evolutionary history did not provide all the opportunities that would have been needed to derive them.

Because they evolved for dealing with a world that is not only complex, but is also constantly changing, these abilities to cope with novel processes (i.e. perceive, represent, and use information about them) at very short notice had to be implemented in architectures that made them

readily available to be invoked on demand in different combinations. I suggest that that fact determines requirements for the design and implementation of visual systems that have not yet been fully articulated. Moreover, the implementation will use mechanisms that have not yet been thought of by neuroscientists, psychologists or AI researchers.

One of the requirements for an organism that may need to monitor, evaluate, modulate and perhaps extend its own mental states and processes (e.g. improving its reasoning, problem-solving, learning, capabilities) is that it should be able to learn not only about the environment but also about its internal states. As with exploration of the environment, this could use a self-organising mechanism that adapts to what it encounters by chunking things and inventing labels for reusable chunks.

This could include labels for aspects of the contents of various sensory manifolds. Because of the manner of their development, such concepts will have a feature referred to as 'causal indexicality', i.e. their intension is intimately connected with their conditions of use. But because they are used for categorising states and processes in virtual machines that are not accessible by anyone else, these concepts will be inherently incommunicable: accounting for one aspect of what people who discuss qualia are trying to say.

When we have designed or discovered appropriate mechanisms for acquiring and using all these different competences, and the kind of architecture required to accommodate them I conjecture that this will explain a wide range of familiar phenomena including the variety of ways in which an individual can experience the world and some of the ways in which things can be experienced as ambiguous, flipping between different interpretations that make use of different competences (or 'techniques').

Some half-baked explorations of these ideas can be found in the html file referenced here

http://www.cs.bham.ac.uk/research/projects/cosy/papers/#dp0601

COSY-DP-0601 Orthogonal Competences Acquired by Altricial Species (Blankets, string, and plywood)

One problem with the theory is that nothing I have learnt about brain mechanisms (on which I am no expert) seems to be capable of explaining how these competences are acquired, stored and recombined on demand.

For example, the kinds of models of neural nets that I am aware of just do not seem to be capable of meeting those requirements, though perhaps networks of networks could? Chemical information processing systems have more of the right features, but would probably be too slow, and could not easily be coupled with the processes that acquire and use the information.

It is possible that there are such mechanisms, but they have not been found because nobody was looking for them. They may be implemented in subtle ways as high level virtual machines on lower level physiological machines that seem to be doing something more mundane or something mysterious.

Note that the recombinability of orthogonal competences seems to require some sort of internal syntax. This could have been a crucial precursor to the development of external social language. It could not be based on human language because the learning and creativity I am talking about occur in prelinguistic children and some other animals.

These ideas have some echoes of global-workspace theory, though I think there are several workspaces of different sorts, supporting different kinds of concurrent processes in the architecture I envisage.

This work is partly inspired by collaboration with Jackie Chappell who studies animal cognition, especially New Caledonian Crows and Parrots/Parakeets.

Machine Consciousness and Machine Ethics

Steve Torrance
Institute for Social and Health Research
Middlesex University
Queensway, Enfield, Middlesex EN3 4SA UK
s.torrance@mdx.ac.uk

Abstract

Questions about the possibility of genuine consciousness existing in future artificial humanoids are closely tied up with ethical considerations. I discuss how the assumed presence or absence of consciousness in artificial persons might make a difference to our ethical attitudes towards them.

Questions about the possibility of genuine consciousness existing in future artificial humanoids are closely tied up with ethical considerations. How might the assumed presence or absence of consciousness in artificial persons make a difference to our ethical attitudes towards them?

For simplicity we will limit ourselves to considering electronic person simulations (EPersons) rather than organic replicates. EPersons could develop to have very rich behavioural and functional properties. Might we expect EPersons to have any ethical responsibilities, or to be subjects of ethical appraisal in any way? And could EPersons have genuine moral interests, or genuine demands on our moral concern? I will consider how the answers to these questions may vary as we consider (a) a condition where EPersons are assumed to possess a form of phenomenal consciousness (and thus can genuinely experience pleasure and suffering) and (b) a condition where they are not assumed to possess such states?

What might be our ethical obligations to such creatures in either the with-consciousness or the without-consciousness conditions? It may be neither rational nor intelligible to bestow moral concern on beings we consider to lack consciousness Conversely, if the behavioural repertoire of EPersons is sufficiently rich and varied, and they enter into a sufficiently wide range of social relations with us, it may be difficult in everyday practice to avoid perceiving or taking EPersons as making legitimate moral claims on us in at least some types of circumstance – even if they are not acknowledged as having phenomenal states.

Could we regard EPersons as having genuine moral responsibility, desert, accountability for their actions or judgments in either of the two conditions? Could they be useful moral advisors? Could they even have a coherent conception of what morality consists of? These questions in part turn on one's view on the role of emotions, and the links between emotions and rationality, in the constitution of a moral agent. It is plausible that our moral conceptions and outlook are derived from our evolutionary inheritance, and are deeply interconnected with a wide range of emotions - anger, envy, compassion, empathy, friendliness, etc.; and that these in turn emanate from biologically-based sentience in natural creatures. EPersons designed around current paradigms of information-processing cognitive architectures, may be incapable of instantiating a deep enough model of emotion and empathetic rationality to support more than a rather impoverished array of moral sentiments at best. This may be true even in the with-consciousness condition; indeed the ability to possess such emotions may be pivotal to the realizability of phenomenal consciousness in EPersons.

I will argue that phenomenal consciousness makes a difference in the cases both of being a genuine bearer of moral responsibilities and of being a worthy recipient of moral treatment. Having fully-fledged or intact phenomenally conscious states is not the sole criterion of moral worth (think of neonates, PVS, dementia); nevertheless the generic property of being the kind of creature that can have phenomenal states is arguably crucial to being a member of the universe of moral concern.

In general we take artefacts to be instruments; so there seems to be an inherently paradoxical quality wrapped up in the idea of extending morality to artificial humanoid beings. However new ways of thinking about moral relationships may forced on us in an era where artificial humanoids live alongside us in significant proportions. A new conception of 'us' may be required.