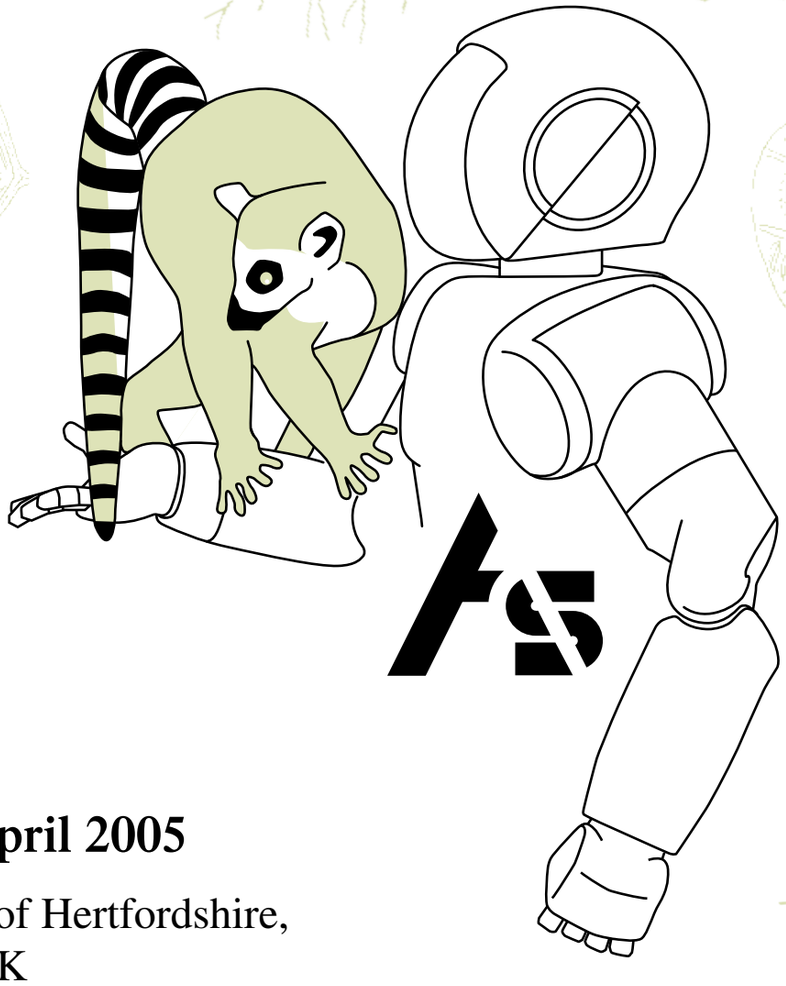
AISB'05: Social Intelligence and Interaction
in Animals, Robots and Agents

Proceedings of the Symposium on Next Generation Approaches to Machine Consciousness

Imagination, Development, Intersubjectivity and Embodiment



12 - 15 April 2005

University of Hertfordshire,
Hatfield, UK

SSAISB 2005 Convention

AISB

UH

EPSRC
Engineering and Physical Sciences
Research Council

AISB'05 Convention

Social Intelligence and Interaction in Animals, Robots and Agents

12-15 April 2005
University of Hertfordshire, Hatfield, UK

Proceedings of the Symposium on

Next Generation approaches to Machine Consciousness:

Imagination, Development, Intersubjectivity, and Embodiment

Published by

The Society for the Study of Artificial Intelligence and the Simulation of Behaviour
www.aisb.org.uk

Printed by

The University of Hertfordshire, Hatfield, AL10 9AB UK
www.herts.ac.uk

Cover Design by Sue Attwood

ISBN 1 902956 46 8

AISB'05 Hosted by

The Adaptive Systems Research Group
adapsys.feis.herts.ac.uk

The AISB'05 Convention is partially supported by:

The proceedings of the ten symposia in the AISB'05 Convention are available from SSAISB:

Second International Symposium on the Emergence and Evolution of Linguistic Communication (EELC'05)
1 902956 40 9

Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action
1 902956 41 7

Third International Symposium on Imitation in Animals and Artifacts
1 902956 42 5

Robotics, Mechatronics and Animatronics in the Creative and Entertainment Industries and Arts
1 902956 43 3

Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction
1 902956 44 1

Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation
1 902956 45 X

Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment
1 902956 46 8

Normative Multi-Agent Systems
1 902956 47 6

Socially Inspired Computing Joint Symposium (Memetic theory in artificial systems & societies, Emerging Artificial Societies, and Engineering with Social Metaphors)
1 902956 48 4

Virtual Social Agents Joint Symposium (Social presence cues for virtual humanoids, Empathic Interaction with Synthetic Characters, Mind-minding Agents)
1 902956 49 2

Table of Contents

The AISB'05 Convention
Social Intelligence and Interaction in Animals, Robots and Agents

Above all, the human animal is social. For an artificially intelligent system, how could it be otherwise?

We stated in our Call for Participation "The AISB'05 convention with the theme *Social Intelligence and Interaction in Animals, Robots and Agents* aims to facilitate the synthesis of new ideas, encourage new insights as well as novel applications, mediate new collaborations, and provide a context for lively and stimulating discussions in this exciting, truly interdisciplinary, and quickly growing research area that touches upon many deep issues regarding the nature of intelligence in human and other animals, and its potential application to robots and other artefacts".

Why is the theme of Social Intelligence and Interaction interesting to an Artificial Intelligence and Robotics community? We know that intelligence in humans and other animals has many facets and is expressed in a variety of ways in how the individual in its lifetime - or a population on an evolutionary timescale - deals with, adapts to, and co-evolves with the environment. Traditionally, social or emotional intelligence have been considered different from a more problem-solving, often called "rational", oriented view of human intelligence. However, more and more evidence from a variety of different research fields highlights the important role of social, emotional intelligence and interaction across all facets of intelligence in humans.

The Convention theme *Social Intelligence and Interaction in Animals, Robots and Agents* reflects a current trend towards increasingly interdisciplinary approaches that are pushing the boundaries of traditional science and are necessary in order to answer deep questions regarding the social nature of intelligence in humans and other animals, as well as to address the challenge of synthesizing computational agents or robotic artifacts that show aspects of biological social intelligence. Exciting new developments are emerging from collaborations among computer scientists, roboticists, psychologists, sociologists, cognitive scientists, primatologists, ethologists and researchers from other disciplines, e.g. leading to increasingly sophisticated simulation models of socially intelligent agents, or to a new generation of robots that are able to learn from and socially interact with each other or with people. Such interdisciplinary work advances our understanding of social intelligence in nature, and leads to new theories, models, architectures and designs in the domain of Artificial Intelligence and other sciences of the artificial.

New advancements in computer and robotic technology facilitate the emergence of multi-modal "natural" interfaces between computers or robots and people, including embodied conversational agents or robotic pets/assistants/companions that we are increasingly sharing our home and work space with. People tend to create certain relationships with such socially intelligent artifacts, and are even willing to accept them as helpers in healthcare, therapy or rehabilitation. Thus, socially intelligent artifacts are becoming part of our lives, including many desirable as well as possibly undesirable effects, and Artificial Intelligence and Cognitive Science research can play an important role in addressing many of the huge scientific challenges involved. Keeping an open mind towards other disciplines, embracing work from a variety of disciplines studying humans as well as non-human animals, might help us to create artifacts that might not only do their job, but that do their job right.

Thus, the convention hopes to provide a home for state-of-the-art research as well as a discussion forum for innovative ideas and approaches, pushing the frontiers of what is possible and/or desirable in this exciting, growing area.

The feedback to the initial Call for Symposia Proposals was overwhelming. Ten symposia were accepted (ranging from one-day to three-day events), organized by UK, European as well as international experts in the field of Social Intelligence and Interaction.

- Second International Symposium on the Emergence and Evolution of Linguistic Communication (EELC'05)
- Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action
- Third International Symposium on Imitation in Animals and Artifacts
- Robotics, Mechatronics and Animatronics in the Creative and Entertainment Industries and Arts
- Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction
- Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation
- Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment
- Normative Multi-Agent Systems
- Socially Inspired Computing Joint Symposium (consisting of three themes: Memetic Theory in Artificial Systems & Societies, Emerging Artificial Societies, and Engineering with Social Metaphors)
- Virtual Social Agents Joint Symposium (consisting of three themes: Social Presence Cues for Virtual Humanoids, Empathic Interaction with Synthetic Characters, Mind-minding Agents)

I would like to thank the symposium organizers for their efforts in helping to put together an excellent scientific programme.

In order to complement the programme, five speakers known for pioneering work relevant to the convention theme accepted invitations to present plenary lectures at the convention: Prof. Nigel Gilbert (University of Surrey, UK), Prof. Hiroshi Ishiguro (Osaka University, Japan), Dr. Alison Jolly (University of Sussex, UK), Prof. Luc Steels (VUB, Belgium and Sony, France), and Prof. Jacqueline Nadel (National Centre of Scientific Research, France).

A number of people and groups helped to make this convention possible. First, I would like to thank SSAISB for the opportunity to host the convention under the special theme of *Social Intelligence and Interaction in Animals, Robots and Agents*. The AISB'05 convention is supported in part by a UK EPSRC grant to Prof. Kerstin Dautenhahn and Prof. C. L. Nehaniv. Further support was provided by Prof. Jill Hewitt and the School of Computer Science, as well as the Adaptive Systems Research Group at University of Hertfordshire. I would like to thank the Convention's Vice Chair Prof. Chrystopher L. Nehaniv for his invaluable continuous support during the planning and organization of the convention. Many thanks to the local organizing committee including Dr. René te Boekhorst, Dr. Lola Cañamero and Dr. Daniel Polani. I would like to single out two people who took over major roles in the local organization: Firstly, Johanna Hunt, Research Assistant in the School of Computer Science, who efficiently dealt primarily with the registration process, the AISB'05 website, and the coordination of ten proceedings. The number of convention registrants as well as different symposia by far exceeded our expectations and made this a major effort. Secondly, Bob Guscott, Research Administrator in the Adaptive Systems Research Group, competently and with great enthusiasm dealt with arrangements ranging from room bookings, catering, the organization of the banquet, and many other important elements in the convention. Thanks to Sue Attwood for the beautiful frontcover design. Also, a number of student helpers supported the convention. A great team made this convention possible!

I wish all participants of the AISB’05 convention an enjoyable and very productive time. On returning home, I hope you will take with you some new ideas or inspirations regarding our common goal of understanding social intelligence, and synthesizing artificially intelligent robots and agents. Progress in the field depends on scientific exchange, dialogue and critical evaluations by our peers and the research community, including senior members as well as students who bring in fresh viewpoints. For social animals such as humans, the construction of scientific knowledge can't be otherwise.

Beppu, Japan.

Dedication:

I am very confident that the future will bring us increasingly many instances of socially intelligent agents. I am similarly confident that we will see more and more socially intelligent robots sharing our lives. However, I would like to dedicate this convention to those people who fight for the survival of socially intelligent animals and their fellow creatures. What would 'life as it could be' be without 'life as we know it'?

Kerstin Dautenhahn

Professor of Artificial Intelligence,
General Chair, AISB'05 Convention *Social Intelligence and Interaction in Animals, Robots and Agents*

University of Hertfordshire
College Lane
Hatfield, Herts, AL10 9AB
United Kingdom

Symposium Preface
Next Generation approaches to Machine Consciousness:
Imagination, Development, Intersubjectivity, and Embodiment

SYMPOSIUM OVERVIEW

A Symposium of *AISB 2005*, University of Hertfordshire, UK
Tuesday 12th and Wednesday 13th April, 2005.

Symposium Co-Chairs: Ron Chrisley (COGS, Sussex), Rob Clowes (COGS, Sussex) and Steve Torrance (COGS, Sussex and ISHR, Middlesex)

Machine Consciousness (MC) concerns itself with the study and creation of artefacts which have mental characteristics typically associated with consciousness such as (self-) awareness, emotion, affect, phenomenal states, imagination, etc. Recently, developments in AI and robotics, especially through the prisms of behavioural and epigenetic robotics, have stressed the embodied, interactive and developmental nature of intelligent agents which are now regarded by many as essential to engineering human-level intelligence.

This symposium explores how new developments in Cognitive Science and associated fields may suggest fruitful new paths for MC research. Contributions to the symposium touch on a number of key themes, including imagination, emotion, development, enactive approaches, heterophenomenology, synthetic phenomenology, second-person approaches, and neurophenomenology. Social, legal and ethical aspects of the development of machine consciousness are also covered. Much of the discussions in the symposium relate to working implementations of robotic or other systems displaying key MC attributes, while a number of contributions are of a more theoretical or synoptic nature. A fuller discussion of the major themes of the symposium and how they are dealt with in the various contributed papers, is given in the first chapter in the proceedings that follow.

The conference chairs would like to thank, for their support in organizing this symposium and in refereeing submissions to the programme, the other members of the programme committee:

Igor Aleksander - Imperial College, UK and COGS Sussex
Giovanna Colombetti - York University, Canada
Rodney Cotterill - Technical University of Denmark
Pentti Haikonen - Cognitive Technology Group, Nokia Research Center, Finland
Germund Hesslow - Department of Physiological Sciences, Sweden
Owen Holland - University of Essex, UK
Takashi Ikegami - University of Tokyo, Japan
Miguel Salichs - University Carlos III Madrid, Spain
Ricardo Sanz - Autonomous Systems Lab, Madrid, Spain
Murray Shanahan - Imperial College, UK
Matthias Scheutz - Artificial Intelligence and Robotics Laboratory, Austria
Jun Tani - Brain Science Institute, Riken, Japan
Tom Ziemke - School of Humanities and Informatics, University of Skövde, Sweden

Next-generation approaches to machine consciousness

Ron Chrisley[*]
ronc@sussex.ac.uk

Rob Clowes[*]
robertc@sussex.ac.uk

Steve Torrance[*†]
stevet@sussex.ac.uk

[*] Centre for Research in Cognitive Science
University of Sussex, Falmer, UK

[†] Institute for Social and Health Research
Middlesex University, Enfield, UK

Abstract

A spate of recent international workshops have demonstrated that machine consciousness is a swiftly emerging field of international presence. Independently, there have been several new developments in cognitive science and consciousness studies concerning the nature of experience and how it may best be investigated. Synthesizing results from embodied AI, phenomenology and hermeneutics in Philosophy, Neuroscience and enactive Psychology (among others), new paradigms for research into natural consciousness that transcend the limited behavioural/cognitive or neural/functional oppositions are being proposed and tested, with encouraging results. This paper gives an overview of some work that attempts to entwine these two strands to see how they might be of mutual benefit to each other.

1 Introduction

The goals of the field of Machine Consciousness are: 1) to create artefacts 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). Machine consciousness symposia in Cold Spring Harbor (2001), Skövde (2001), Memphis (2002), Birmingham (2003), Turin (2003) and Antwerp (2004) have demonstrated that this is a swiftly emerging field of international presence.

Independently, there have been several new developments in cognitive science and consciousness studies concerning the nature of experience and how it may best be investigated. Synthesizing results from embodied AI, phenomenology and hermeneutics in Philosophy, Neuroscience and enactive Psychology (among others), new paradigms for research into natural consciousness that transcend the limited behavioural/cognitive or neural/functional oppositions are being proposed and tested, with encouraging results.

Next-generation approaches to machine consciousness attempt to entwine these two strands to see how they might be of mutual benefit to each other. A guiding principle behind this union is that advances in consciousness research can guide efforts into building conscious systems. But equally, there is a belief that the converse is true: The insights gained from attempting to build embodied, experiencing agents can provide important feedback to the various disciplines of consciousness studies. At the very least, the difficulties we encounter in our attempts to build systems which instantiate or model cognitive phenomena can point out where our current theories are incomplete, inadequate or incorrect.

The symposium entitled "Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment" (part of AISB 2005) brings together active researchers in this area and provides a forum in which their work may be compared, contrasted, evaluated and discussed. This paper uses the work being presented at that symposium as a framework around which to organise a survey of some of the key strands in contemporary work in machine consciousness.

The term "next-generation" may be something of a misnomer, since there is no clear consensus as to what constituted "first-generation" approaches to machine consciousness; certainly no attempt will be made here to provide a scholarly account of earlier approaches in this area. Nevertheless, we feel that the application of recent advances in our understanding of consciousness to the construction of working systems constitutes a major milestone on the way to achieving machine consciousness.

After a general discussion of the enterprise of machine consciousness in section 2, sections 3-10 examine what we believe to be eight key areas of consciousness studies that are best placed to help make progress on machine consciousness: work on imagination, emotion (and feelings of emotion), development and self-creation, enactivism, heterophenomenology, synthetic phenomenology, second-person approaches and neurophenomenology, and ethical and legal issues.

2 Machine consciousness: The very idea

As Torrance (2005, this volume) points out, one can revisit Searle's (1980) old distinction between weak and strong AI and similarly differentiate between weak and strong forms of machine consciousness. The first attempts merely to model conscious states in an artificial system, without any ambitions of actually replicating consciousness in that system. Strong machine consciousness goes further, and seeks to create artificial systems with experiential states themselves.

One might think it folly to engage in machine consciousness research (especially of the strong variety), given the opposition that confronts it on both sides. On the one hand, there are some popular arguments (e.g. Jackson, 1982) against physicalist accounts of consciousness, which claim that some form of dualism is the case. On the other hand, there are the arch-physicalists, who define consciousness in such a bio-centric manner that no non-biological system, such as the ones with which the field of machine consciousness typically concerns itself, could ever be conscious. Machine consciousness appears trapped between these two extremes; surely one or the other of them must be correct, and yet both rule out the possibility of conscious machines.

Although this is not the proper place for a full reply, a quick response can be made, since it works equally well against both lines of attack. Many in the machine consciousness community take what has been termed (since the Birmingham meeting) an "engineering" approach. Rather than claiming to have a solution to the "other minds" problem that would let them know definitively whether or not this or that artefact is or could be conscious, these researchers are more pragmatic. Modifying a criterion from the field of artificial intelligence, they will have considered their goal accomplished if they design and construct a system that does the kinds of things that, when done by a human, requires consciousness. It is sufficient for them to produce a system that behaves in such a way that, if it were an organism, we would assume that it is aware.

Many researchers in the area find the axioms provided by Aleksander and Dunmall (2003) to be of assistance in guiding their research. More generally, it is thought that the goal of the field will have been achieved if one can impart to a robot some combination of features, possibly including some of the following:

- Autonomy
- Adaptivity/advanced learning capacities
- Emotion/affect
- Responsibility (or being something to which we are responsible)
- Intelligence
- Authenticity (own world view and goals)
- Ability to integrate information from different sources/modalities
- Vivid/meaningful sensation/perception
- Ability to act in the world
- Ability to simulate/imagine/plan
- Ability to represent its own states
- Attentional capacities
- A belief that it is conscious/an ability to give phenomonological reports

Certainly most in the field would consider their primary goal achieved if they could build a system which had all of these features, even if philosophical doubts as to whether the system is "really" conscious might remain.

But even on that point, there is room for optimism. If Sloman and Chrisley (2003) are right, then current philosophical puzzles concerning how we could ever know a machine is conscious (a product of the apparent possibility of "zombies" (Chalmers, 1996)) might be features of our current, inchoate concept of consciousness. Perhaps we can't get to the successor concepts of consciousness that will solve these problems through armchair theorizing alone. But if we design, build and interact with artefacts with some of the properties listed above, that might be enough to cause our concept of consciousness to evolve until we see that no, it isn't possible for a system to have *this* architecture and not be aware. Zombies may seem possible now, but the kind of research surveyed in this paper might someday reveal that they actually are not possible

2 Imagination

A key finding of the Birmingham and Turin meetings was the existence of a common theme in much of the recent work in machine consciousness: The imagination or simulation approach. The basic idea is of an ability to predict, given the current sensory input, the future sensory input one would receive if one were to make a particular motor

response. If this predicted sensory input is used as the "current" sensory input for an iterated application of the predictive process, one can anticipate the sensory input one would receive if one made a second motor response after the first, and so on. This allows entire sequences of behaviours, with the corresponding sequences of sensations that would occur during that behavioural sequence, to be "imagined".

The idea of using a simple recurrent network to give a robot this kind of imaginative capacity is not particularly new (see, e.g., Chrisley, 1990). But from the start it was acknowledged that imaginative capacities that dealt only in the lowest levels of sensory and motor encodings would be extremely limited. The work of Stening, Jacobsson and Ziemke (2005, this volume) is therefore a welcome development in this area of research. Not only do they incorporate an abstraction mechanism that allows their robot to imagine at a higher level of "conceptualisation" than the lowest sensory and motor levels; they also provide an inversion mechanism so that the imagined abstract states can be converted into expected sensory-motor states. A future extension of this work might be to have both low-level and abstract-level imaginative capacities working simultaneously, so that expected low-level sensations can be fed into the abstraction mechanism to yield a second route to abstract expectations. Actual abstract expectations might be some kind of average between the "abstract-then-imagine" expectations and the "imagine then abstract" expectations outlined here.

Shanahan (2005, this volume) illustrates the imagination approach very well. He reports on a new kind of design for robot architectures that incorporates two linked action-generation systems, a first-order reactive system and a higher-order one which introduces off-line 'imaginative' rehearsals of action alternatives in a way that modifies the saliency levels made available to the first-order system. The resulting architecture incorporates various key features of mammalian brains. The function of imaginative rehearsal plays a key role in the model of consciousness offered by Shanahan. The model provides, in his view, a useful approximation to the role played by consciousness in real agents.

It is hard to say exactly what it is about imaginative processes that makes some researchers take imagination to be essential to consciousness. For some, such as Haikonen (2005, this volume) and Stening et al. (2005, this volume, following Hesslow (1994)), consciousness consists in having an inner life or inner world, and it seems more plausible to say an artificial system has such if one can identify states of the system that are of the same format as perceptual states, but which correspond to anticipated rather than actual sensory input. The imagination approach, with its extension of perceptual processes to cognition as a whole can be seen as a new kind of empiricism. Yet, as Haikonen points out, imagination allows one to transcend perception, in that one's behaviour may sometimes be driven by internal (albeit pseudo-perceptual) processes rather than the current sensory input. A striking feature of his model is its attempt to go beyond the simplest models of artificial imagination, by integrating it with elements such as attention and decision-making.

Stuart (2005, this volume) also addresses the relation of imagination to consciousness, but in a rather different way. She suggests that Kant's transcendental philosophy prefigures a variety of recent studies of artificial agency and consciousness – particularly the work of Cotterill, Sloman, Aleksander, Bowling and others. Her focus is on Kant's treatment of the problem of how it is we can take the order of our experiences as belonging to a unified 'I' – a precursor of the contemporary binding problem. Kant's solution is complex but, as Stuart shows, a central strand appeals to the imagination, specifically the cognitive or productive imagination that (working with the senses and the understanding) enables us to treat each of our experiences as modifications of he same mind; as linked in consecutive, associative patterns; and as similar or different from preceding experiences. Kant distinguishes between productive and reproductive imagination: the first is essential for any thought and necessary for the constitution of self-consciousness; the second is the ability to bring to mind things that are not wholly present. No doubt both types of imagination are required within an adequate model of consciousness.

One of the first intended applications of imagination in robots was planning (Chrisley, 1990; Stein, 1995). Chella, Frixione, and Gaglio (2005, this volume) combine this idea with a linguistic abstraction capability to allow for grounded planning for linguistically-specified goals. It seems that their system could be extended to also allow for linguistically-specified environmental information to play a role in the planning process.

3 Emotion (& feelings of emotion)

In recent years the seeming antithetical study of emotions in machine systems has started to be treated seriously (Picard, 1997). In work on machine consciousness, Aleksander, Lahnstein, & Lee argue, one should be "suspicious of the consciousness of a machine were it not to have mechanisms that play the role of emotions" (2005, this volume). They maintain that valenced evaluation of the state of the

organism, both actual and projected, are central to the long-term viability of, and the development of the capacities of, the organism. Some researchers in machine consciousness seek to develop this idea with reference to the ideas of Antonio Damasio.

According to Damasio, emotion is not only central to reasoning (Damasio, 1994) but to the generation of what he calls core consciousness (Damasio, 2000). On Damasio's account, core consciousness emerges for an organism as it becomes able to detect that its core body state has been changed by some incoming stimulus. The reactive component of the organism's neural representation of such a stimulus is conceptualized as an emotion. Bosse, Jonker, & Treur (2005, this volume) formalize this theory into a model expressing temporal and causal dependencies using their Temporal Trace Language (Jonker & Treur, 2002). Their formal model also predicts the possibility of "false core consciousness", where an effect is attributed to the wrong body stimulus.

Aleksander et al. (2005, this volume) build upon Damasio's model in order to understand a key point of discussion in the (natural) consciousness literature, that is, accounting for the reality or otherwise of "the will". Since the publication of Libet et al's (1983) finding that a neo-cortical readiness potential seemed to precede the ability of a subject to attest to willed action, the folk conception of volitional action has been called into question. One radical sceptic (Wegner, 2002) has recently argued that Libet's findings should be interpreted as showing that an unconscious cortical event controls both the "willed" action itself and the conscious sensation of control. By this reasoning, "the will" as currently conceptualised by the folk is simply an epiphenomenal shadow or illusion. Aleksander et al. (2005, this volume) instead see Libet's experiment as having taken volition out of its normal emotional envelope. By developing a model of how such volitions are typically generated within a framework of ongoing affective evaluations, the authors show that Libet's paradigm is actually an atypical example of willing where the will is relegated to *when* and not *what*. If volition is examined in its typical and proper emotional context, they argue, it approximates much more closely the way it is seen by the folk.

No doubt the next round of machine consciousness research will pay more attention to emotion and affect. In a rather plausible Humean way, Haikonen (2005, this volume) contends that an "emotional value system", or at least some affective distinction between pleasant and unpleasant, is required for decision making (in his case, via an imaginative system). Stening, Jacobson and Ziemke (2005, this volume) make a similar point, noting that future development of their work should allow the robot's needs to play a role in motivating and guiding its action, abstraction and imagination.

4 Development and self-creation

Development has for a long time been argued to be a crucial component in the understanding of consciousness. Vygotsky (1986), for one, pointed out this link by attempting to show how consciousness depended upon intersubjective social interactions. Although the connections between these areas remain largely unaddressed, work on the development of intersubjectivity (Trevarthen, 1994) may point the way forward; indeed, this work is starting to be taken very seriously in the related field of epigenetic robotics.

Of course, development appears to depend not just on external scaffolds but also on the developing bodily and situational substrate, and it is the attempt to understand the relation between these that has been fundamental to the concerns of epigenetic roboticists. A series of annual conferences in this field (starting in 2001) has focused on the question of how a robotic system, through extensive interaction with its environment, can transform itself from a being a purely reactive system into a fully intentional one. One idea is that this can happen only if an agent undergoes a prolonged developmental period (Zlatev, 1999). Central issues in the development of agents are the distinctions between self and other, body and environment, sense and action.

Such questions should also be germane to work on machine consciousness, not least because some would be unwilling to treat as conscious any system that was incapable of undergoing a process of ontogenesis – although this is controversial. This of course throws open the question of what forms of development might evince consciousness. One possible focus is the development of self.

As having a self – generally one to a body – seems to be typical of the type of consciousness we best know, i.e. our own, systems that attempt to explore the development of self should be of special interest. Many accounts of the self stress that a sense of self is not pre-given to an agent or merely represented internally, but is developed and maintained out of the sensorimotor flows in which the agent participates (Butterworth, 1998). Although there is considerable controversy over what is pre-given and what is developed (see for instance Gallagher & Meltzoff, 1996), the flexibility of the nature of the body image in higher animals now has extensive experimental demonstration (Ramachandran & Blakeslee, 1998). Iizuka & Ikegami (2005, this volume) argue that "body image and ownership" – concepts that seem closely related

to the idea of the self – cannot be derived from static sense data alone. They argue that the self depends on and must be understood in terms of the emergence of the distinction between self and world. Inspired by Gibson's (1962) cookie-cutter experiment, they discuss a simulated agent that develops distinctions between 'sensor' and 'motor' through interactions with its world. They argue that the sort of active perception system here developed can help us understand the emergence of a self in a way which is precluded by the prior specification of sense and motor.

5 Enactive approaches

Enactive approaches to cognitive science have become popular of late. Enactivism was first formulated as an attempt to move beyond cognitive science methods dominated by cognitivist and connectionist paradigms (Varela, Thompson and Rosch, 1991). Strong emphasis was laid on linking cognitive science with insights from the hermeneutic phenomenology of Husserl and Merleau-Ponty, and in particular stressing the sensorimotor embodiment of an experiencing agent in a world, "enacting" that world and her own self in relation to the world.

There are a number of strands to the enactive approach. One focuses on perceptual experience, arguing that it consists in the exercise of the mastery of sensorimotor contingencies, and that awareness consists in the application of this mastery to a reasoning process (e.g. O'Regan and Noë, 2001; Noë, 2002). This view contrasts with the conventional view of perceptual awareness, according to which experience consists in sensory inputs generating internal, neurally-encoded representations of an external environment. For the enactive approach, perceptual consciousness has relatively little to do with intenal structures in the brain, and much more to do with ongoing sensorimotor and bodily interactions with the environment.

In this respect the enactive approach contrasts quite strongly with at least some variants of imagination-based approaches to modelling consciousness – for example that of Shanahan, who puts considerable emphasis on providing an architecture that reproduces detailed structures of the brain. Another prominently neurophysiologically-based approach to consciousness which, however, also lays great stress on embodiment and sensorimotor fusion in a way that is close to the enactive approach, is to be found in Cotterill (1998). Stuart (2005, this volume) considers Cotterill's approach in some detail, putting it into the context of Kant's debate with Hume over the nature of the unity of self-consciousness. She points out that an adequate account of the unity of the experiencing and active "I" must necessarily be strongly embodied, and thus her approach is also close to that of the enactive viewpoint.

Both Haikonen (2005, this volume) and Ikegami (2005, this volume) take there to be a fundamental connection between consciousness and enactive perception, at least in the Gibsonian sense that perceptual experience is not the passive reception of sensory inputs, but an exploratory interplay between the internal states of the agent and the external world. Haikonen points out that if perceptual experience consists in active exploration of sensory-motor contingencies, then it makes sense that imaginary or inner experience consists in the exploration of the interdependence between hypothetical motor commands and the anticipated sensory states which result. Ikegami's focus, however, is on exploration in the real world rather than in some inner simulation. He attempts to model this with an agent whose chaotic dynamics are such that the agent, he says, is not simply responding to the stimuli at any one time, but to a more abstract entity: the time structure of the stimuli. However, it is not yet clear whether such a dynamics has the property which Ikegami seeks: that of being able to specify what is characteristic of conscious states (or even living states. For Ikegami, life seems to be a prerequisite for consciousness, a contentious view in the machine consciousness community).

If the first-strand enactivists are right, then perceptual experience consists in the exercise of the mastery of sensory-motor contingencies, and that awareness consists in the application of this mastery to a reasoning process. In that case, the central goals for machine consciousness research would be a) establishing clear criteria for when a robotic system possesses such mastery and b) building robots which meet these criteria in a way which allows said mastery to play a crucial role in their deliberations.

A second strand in enactivism goes back to Varela's earlier work, with Maturana, on autopoiesis (Maturana and Varela, 1987). Autopoiesis is the process whereby an organism continually recreates itself in relation to its environment, through a process of internal self-regulation and the maintenance of a semi-permeable boundary through which matter or energy can be exchanged. There has been some interesting recent philosophical work exploring the implications of autopoiesis in ways that help to understand the nature of consciousness. Torrance (2005, this volume) discusses the significance of some of this work (e.g. Hanna and Thompson, 2003; Weber and Varela, 2002) in offering a new departure for machine consciousness. He considers an impasse over the 'explanatory gap'

which is seen by many as blocking physicalistic attempts to explain the nature of phenomenal consciousness Torrance suggests that there is a defective concept of consciousness underlying this gap – 'thin phenomenality' as he calls it – which is also shared by many of those who think the gap can be bridged, including many machine consciousness researchers. An alternative, 'thick' conception of phenomenality is proposed, which takes ideas of autopoiesis, lived embodiment and other related ideas as its starting point.

6 Heterophenomenology

It seems undeniable that phenomenological reports are a valuable source of data concerning consciousness, and yet a scientific theory of consciousness must be sensitive to the possibility that subjects may be mistaken in their sincere avowals concerning experience. Dennett (1991; 2003) outlines a way to avoid the pitfalls of naïve or folk conceptions of consciousness without discounting phenomenological reports altogether: Heterophenomenology. Adopting this methodology with respect to machine consciousness seems promising, but poses difficult questions. For example, since linguistic phenomenological reports play such an important role in this approach, what kinds of communicative or linguistic abilities need a robot possess in order to allow the direct application of heterophenomenology?

Modelling and robotic work such as the Adaptive Language Games project of Steels (1998) and his collaborators has provided one way into understanding how mechanisms for grounding communicative symbols in perceptual abilities might be effected. In a recent extension to this work Steels (2003) has argued that a variation on the adaptive language games model can be used to help understand the inner-voice which is thought to be the constant accompaniment of much human conscious thinking (Hurlburt, 1990). In Steels' model, agents pre-check the interpretability of a putative sentence by feeding back the output from their production systems into their interpretation systems. It is argued that this "re-entrance" of linguistic information where an agent checks an utterance by projecting it back onto itself, explains the functional system underlying the phenomenology of inner speech.

Other work on linking cognitive and linguistic functions can be found in Sugita & Tani's (2002) report on their work with a mobile robot, where the robot comes to associate action categories and linguistic labels. Chella, Frixione, & Gaglio (2005, this volume) report on their work on Cicerobot, where a comparable approach is taken but on a more sophisticated mobile robot. In this research the authors have built a robot capable of vision and action which has an architecture based on linguistic, conceptual and sub-conceptual capacities. Cicerobot's architecture, however, is based on a three-layer model composed of a "subconceptual area… concerned with the processing of data coming from the robot sensors…, [a] linguistic area of representation and processing… based on a semantic network formalism… [and a] conceptual area intermediate between the subconceptual and the linguistic areas." Cicerobot's linguistic and subconceptual areas are used in behavioural planning and affective evaluations, and these different representational levels are mediated by the 'conceptual area'. The authors make use of conceptual space theory (Gärdenfors, 2000) to "provide a principled way for relating high level, linguistic formalisms with low level, unstructured representation of data." It would be interesting to see how this work might be developed to support the sort of phenomenological reports required for heterophenomenology.

However, the generation of narratives which would serve the role assigned for them by Dennett would seem to require the involvement of language in the ongoing activity of the agent in a way which would need to go somewhere beyond the labelling by the agent of its environment or even a role in planning (Clowes, 2003). Whether this work can provide a sufficient underpinning for machine heterophenomenology remains to be shown, but we are starting to have a better range of possible scenarios to consider.

Another direction in which to pursue this approach would be to try to make sense of infra-linguistic forms of phenomenological "reports". It seems possible at least in principle that a system incapable of using language might nevertheless attempt to represent its internal states *as* phenomenological states. Indeed, one might think such self-modelling might be a crucial component in explaining even the phenomenological reports of linguistic creatures. As Sloman and Chrisley (2003) point out, explaining why a system finds it useful to think (or speak) of itself as having qualia might go a long way to explaining the having of qualia itself.

7 Synthetic phenomenology

A science of consciousness, be it of natural or artificial agents, requires some ability to specify and refer to subjective, fine-grained experiential states, which, by their very nature, elude linguistic expression. One idea is that the states of artefacts-in-an-environment might themselves serve as ways of specifying the conscious states that they embody

(Chrisley, 1995). The sub-field of synthetic phenomenology aims to investigate this idea by, e.g., constructing means of visualizing or otherwise communicating the (actual, or modelled) experiential states of robots.

It has been known for some time that capturing the spatial content of experience is particularly problematic. Previous attempts to do so have simplistically plotted the robot's actual or imagined sensations on a map of objective space, even when the robot had no understanding of the connection between the spatial content of the sensors and movement, and even when the non-objective, non-systematic spatial representations of the robot were explicitly the topic of investigation (e.g., Chrisley, 1993). Another difficult area for synthetic phenomenology, discussed at the Antwerp meeting, is the specification of the content of experience which is more abstract or conceptualized than the lowest level of sensory and motor signals.

Stening et al. (2005, this volume) manage to make headway on both problems with a single solution: de-abstraction, or "inversion". Their initial representations of the abstract aspects of their robot's experience suffer from the usual problems: They are located on a map of objective space, and their forms (e.g., their gray-scale ordering, their circular shape) do not carry any content for us that is related to the contents of the robot which contain those abstractions. But their later representations of the robot's experience do much better: By inverting the sequence of abstractions into sensorimotor combinations, they are able to reveal the spatial relational structure of the robot's experience. The inversion, by reducing abstractions back to the sensorimotor level, allows a more helpful depiction of the content of the robot's experience. (Compare the "Anchored" c-knoxels in Chella, et al (2005, this volume)). But as it stands, the method may be too reductionistic on this second point. It seems desirable to have some way to distinguish notationally the experiences of a robot that produces those sensorimotor expectations directly, from that of a robot that has those same expectations as a result of an abstraction and de-abstraction process.

Stening, et al's "inversion" method allows them to compare the phenomenological world of a three-category robot to that of a five-category robot in a way that reveals the latter to be much more akin to how we experience the objective structure of space. But they also point out that the inversion method allows one to make relative comparisons that are essential for gauging the imaginative abilities of systems that have experiences that are fundamentally different from our own. Specifically, their method allows one to see that the three-concept robot's imagined world faithfully reconstructs its perceived world, even though both are radically different from how we would experience that space. A less subtle form of synthetic phenomenology, that merely focussed on the three-concept robot's inability to reconstruct *our* experience of the space, would have been unable to identify these successful aspects of the robot's imaginative capacities.

Synthetic phenomenology is also the focus of Gamez (2005, this volume). His review of the recent consciousness literature leads him to a position close to that of Prinz (2003) that "no test can separate out necessary and sufficient correlates or causes of consciousness. We can vary the ways in which the global functions of the brain are implemented in a vast number of ways, but since these will always lead to the same behavioural output… [and even to the same phenomenal experience from the first person perspective], any impact of these changes on consciousness cannot be measured and we will never know for certain whether a functionally… identical robot has conscious states at all." However, Gamez does not think this stance prohibits the development of a synthetic phenomenology. The paper develops an ordinal probability scale which is designed to be used in assessment of the possibility that our artificial creations might have consciousness. Gamez's contention is that the development of the field will eventually necessitate the research community and society at large to require just such a scale which will be of use in judging the development of the field both in its own terms, and for ethical purposes. For example, creating machines even with the strong likelihood of the capability of suffering might be intrinsically ethically problematic (but see section 9, below), and so it would be of great ethical import to be able to have some principled manner of assessment beyond personal intuition.

Having said that, Gamez's ordinal probability scale proposes formalizing our intuitions in a manner perhaps quite related to the axiomatic approach of Aleksander & Dunmall (2003). However, unlike those authors, Gamez is, as said before, skeptical about the possibility of developing strong axioms. Instead, building on Harnad's (1994) extension of the Turing test, Gamez proposes a metric for consciousness based on similarity to ourselves. The scale is thus strongly anthropocentric and by necessity will have difficulty accounting for other possible kinds of consciousness. Using the scale Gamez analyses several existing systems: Lucy (Grand, 2003); Demarse, Wagenaar, Blau, & Potter's Neurally Controlled Animat (2001); IDA (Franklin, Keleman, & McCauley, 1998); and, after Block (1978), a fictional functional system implemented by the population of China, in order to assess their respective likelihoods of being conscious.

8 Second-person approaches and neurophenomenology

The term 'neurophenomenology', (originating, like the 'enactive' approach, with Varela (1996; see also Thompson, Lutz and Cosmelli, 2005)), denotes the fusion of hermeneutic philosophy with rigorous empirical methods in neuroscience. A key element in neurophenomenology is the use of systematic techniques to enable phenomenologically trained subjects to give precise first-person accounts of features of their experiences. Second-person approaches, also favoured by Varela and others, stress empathetic interaction as a way of understanding consciousness. Social interaction – especially the notion that human consciousness develops from and is grounded in intersubjective processes – has been fundamental to the growth of first- and second-person studies in consciousness (Varela and Shear, 1999; Thompson, 2001). The sophisticated, interactive protocols being developed in Neurophenomenology may prove to be a source of data and design intuitions for the construction of systems that merit the attribution of phenomenological states. Since theorists are themselves social subjects, in giving an account of experience one cannot ignore the intersubjective relationships between theorist and subject (or robot).

The work of Nomura, Takaishi and Hashido (2005, this volume) has some relevance to this theme. They explore how virtual and robotic agents displaying many characteristics of consciousness (e.g. affective, empathic interactions) are perceived by participants in social settings such as psychotherapy and healthcare. The primary interest of Nomura and colleagues is in the psychological and sociological features of such applications. Their use of the term 'machine consciousness' stands somewhat in contrast to the more tentative use of those who regard machine consciousness somewhat as a 'holy grail' to be arrived at possibly only in the remote future. For Nomura et al., any system which is taken by (albeit naïve) users as possessing characteristics associated with conscious agents, may be taken to exemplify "machine consciousness" – so that even simple Eliza-style systems may display a schematic variety of that property. Even if "genuinely" or "literally" conscious machines lie in the realm of "science fiction" (as Shanahan would have it), the proliferation of computational agents displaying complex conscious-like characteristics that are taken by many to be signs of real consciousness may soon be a sociological fact. The social ramifications of the mass arrival of such pseudo-conscious agents are likely to extend over many other aspects of society than just therapeutic applications.

9 Ethical and legal issues

Some would argue that machine consciousness (unlike "mere" machine intelligence) has an inherently *ethical* dimension. A genuinely conscious machine (rather than one which merely shows outward signs or internal organizational features of consciousness) would perforce be capable of enjoyment, suffering, etc., and thus apparently be a genuine ethical subject (Torrance, 2000a; 2000b). If this is so, then the ethical dimensions of machine consciousness research can not really be treated as something external to the research enterprise. Rather, as we build increasingly complex artefacts in order to understand consciousness, normative concerns become essential, both to our understanding of the constitution of subjectivity, and to our appreciation of, and actions towards, the artefacts we create. These and other issues concerning the ethical import of machine consciousness are discussed by Torrance (2005, this volume).

Torrance cites the warning, expressed by Thomas Metzinger (2003), that, since being a conscious creature necessarily involves the possibility of great suffering, the development of artificially conscious creatures is perhaps an activity which we are morally obliged not to even start on. This may be a rather overzealous prohibition – our children will probably suffer to some degree or other during their lives, but we are surely not for that reason morally forbidden from procreating. But the point does lay down a strong challenge to strong machine consciousness researchers to become more aware of the ethical dimensions of their activities. The artificial consciousnesses we create won't be like our human children, and their differences from us may be profound and unpredictable.

Quite apart from the difficult moral and social questions raised by the machine consciousness enterprise there are also the legal questions. Calverley (2005, this volume) considers some of the relevant foundational issues in jurisprudence. He particularly considers the implications of debates between supporters of natural and positive conceptions of law, for the possible future emergence of artificial autonomous agents displaying features of consciousness. What extensions should be made within existing human-based legal – and moral – frameworks to properly take account of such agents? It seems clear that it will be necessary to clarify what kinds of legal *responsibilities* future autonomous machine consciousness agents might have, and also what legal *rights* we should accord them – what responsibilities we may have towards them. Calverley considers such questions in some depth,

taking as his point of departure discussions that have already been initiated between cognitive scientists and lawmakers in the United States.

Acknowledgements

The authors would like to thank everyone who has assisted in the preparation of this work, including the E-Intentionality research group in COGS at the University of Sussex, the participants of the Skövde, Birmingham, Turin and Antwerp machine consciousness meetings, and the programme committee and authors for the AISB05 Symposium on Next Generation Approaches to Machine Consciousness.

References

Aleksander, I., & Dunmall, B. (2003). Axioms and Tests for the Presence of Minimal Consciousness in Agents. *Journal of Consciousness Studies*, 10, 7-18.

Aleksander, I., Lahnstein, M., & Lee, R. (2005). Will and Emotions: A Machine Model that Shuns Illusions. 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.*

Block, N. (1978). Troubles with Functionalism. In C. Wade Savage (Ed.), *Minnesota Studies in the Philosophy of Science*, (Vol. IX). Minneapolis: University of Minnesota Press.

Bosse, T., Jonker, C. M., & Treur, J. (2005). Simulation and Representation of Body, emotion and Core Consciousness. In Chrisley, R., Clowes, R., and Torrance, S. (Eds.) *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Butterworth, G. (1998). A developmental-ecological perspective on Strawson's 'The Self'. In S. Gallagher & J. Shear (Eds.), *The Self.*

Calverley, D. J. (2005). Towards a Method for Determining the Legal Status of a Conscious Machine. 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.* (pp. 75-84).

Chalmers, D. (1996) *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Chella, A., Frixione, M., & Gaglio, S. (2005). Planning by imagination in Cicerobot, a robot for museum tours. In Chrisley, R., Clowes, R., and Torrance, S. (Eds.) *Proceedings of the AISB05 Symposium on Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment.*

Chrisley, R. (1990). Cognitive map construction and use: A parallel distributed processing approach," in Touretzky, D., Elman, J., Hinton, G., and Sejnowski, T. (Eds.) *Connectionist Models: Proceedings of the 1990 Summer School.* San Mateo, CA: Morgan Kaufman. pp 287-302.

Chrisley, R. (1993). Connectionism, cognitive maps, and the development of objectivity. *Artificial Intelligence Review* **7**, pp 329-354.

Chrisley, R. (1995). Non-conceptual content and robotics: Taking embodiment seriously. In Ford, K., Glymour, C. and Hayes, P. (Eds.) *Android Epistemology*. Cambridge: AAAI/MIT Press, pp 141-166.

Clowes, R. W. (2003). Action Oriented Adaptive Language Games. Presented at the Third International Workshop on Epigenetic Robotics, Boston.

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

Damasio, A. R. (1994). *Descartes' Error: Emotion, Reason and the Human Brain*. London: Papermac.

Damasio, A. R. (2000). *The Feeling of What Happens: Body, emotion and the making of consciousness*. Vintage.

Demarse, S., Wagenaar, A., Blau, A., & Potter, A. M. (2001). The neurally Controlled Animat: Biological Brains Acting with Simulated Bodies. *Autonomous Robots*, 11(3), 305-310.

Dennett, D. C. (1991). *Consciousness Explained*. Boston: Little, Brown..

Dennett, D. C. (2003). Who's On First? Heterophenomenology Explained. *Journal of Consciousness Studies*, 10, 19-30.

Franklin, S., Keleman, A., & McCauley, L. (1998). IDA: A Cognitive Agent Architecture. *IEEE*

International Conference on Systems, Man and Cybernetics, 3, 2646-2651.

Gallagher, S., & Meltzoff, A. (1996). The Earliest Sense of Self and Others: Merleau-Ponty and Recent Developmental Studies. *Philosophical Psychology*, 9, 213-236.

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

Gärdenfors, P. (2000). *Conceptual Spaces the Geometry of Thought*. London, England: MIT Press.

Gibson, J. J. (1962). Observations on active touch. *Psychological Review* (69), 477-491.

Grand, S. (2003). *Growing up with Lucy*. London: Weidendfield & Nicolson.

Haikonen, P. (2005). 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.*

Hanna, R. and Thompson, E. 2003. The mind-body-body problem. Theoria et Historia Scientiarum: *International Journal for Interdisciplinary Studies* 7.

Harnad, S. (1994). Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life. In *Artificial Life 1* (Vol. 3).

Hesslow, G. (1994). Will neuroscience explain consciousness? *Journal of Theoretical Biology* **171**, 29-39.

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

Iizuka, H., & Ikegami, T. (2005). Emergence of Body Image and Dichotomy of Sensory and Motor Activity. 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.*

Ikegami, T. (2005). 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.*

Jackson, F. (1982) Epiphenomenal qualia. *Philosophical Quarterly* **32**, pp. 127-36.

Jonker, C. M., & Treur, J. (2002). Compositional Verification of Multi-Agent Systems: A Formal Analysis of Pro-Activeness and Reactiveness. *International Journal of Cooperative Information Systems*, 11, 51-92.

Libet, B., Gleason, C., Wright, E., & Pearl, D. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): The unconscious imitation of free voluntary activity. *Brain*, 106, 623-642.

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

Metzinger, T. (2003) *Being No One: The Self-model Theory of Subjectivity.* Cambridge, MA: MIT Press.

Noë, A. (2002). Is the visual world a Grand Illusion? *Journal of Consciousness Studies.* 9 (5/6), 1-12.

Nomura, T., Takaishi, K., & Hashido, T. (2005). Considerations of Machine Consciousness in the Context of Mental Therapy from Psychological and Sociological Perspectives. 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.*

O'Regan, J.K. and Noë, A. (2001) A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences,* 24 (5). 883-917

Picard, R. (1997). *Affective Computing*. Cambridge, Mass.: MIT Press.

Prinz, J. J. (2003). Level-Headed Mysterianism and Artificial Experience. In O. Holland (Ed.), *Machine Consciousness*. Exeter: Imprint Academic.

Ramachandran, V. S., & Blakeslee, S. (1998*). Phantoms in the brain*. New York: Harper Collins.

Searle (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences* **3**, pp. 417-24.

Shanahan, M. (2005). 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.*

Sloman, A. and Chrisley, R. (2003). Virtual machines and consciousness. *Journal of Consciousness Studies* **10**:4-5, pp 133-172

Steels, L. (1998). Structural Coupling of Cognitive Memories Through Adaptive Language Games. In R. Pfeifer, B. Blumberg, J.-A. Meyer & S. Wilson (Eds*.), Animals to Animats 5: Proceedings of SAB 98* (pp. 263--269). Edinburgh: The MIT Press.

Stein, L. (1995). Imagination and situated cognition. In Ford, K., Glymour, C. and Hayes, P. (Eds.) *Android Epistemology*. Cambridge: AAAI/MIT Press.

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

Stening, J., Jacobsson, H., & Ziemke, T. (2005). Imagination and Abstraction of Sensorimotor Flow: Towards a Robot Model. 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.*

Stuart, S. (2005). 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.*

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)*.

Thompson, E. ed. (2001). *Between ourselves. Second-person approaches to the study of consciousness.* Thorverton: Imprint Academic.

Thompson, E., Lutz, A. & Cosmelli, D. (2005) Neurophenomenology: An Introduction for Neurophilosophers in: A. Brook & K. Akins (Eds.) *Cognition and the Brain: The Philosophy and Neuroscience Movement.* Cambridge University Press.

Torrance, S (2000a) Ethics, Mind and Artifice, in R. Chrisley (Ed.) *Critical Concepts in Cognitive Science,* vol. 4. London: Routledge, 2000. (Reprinted from K.S.Gill (ed) *AI for Society.* Chichester: John Wiley, pp 55-72., 1986.)

Torrance, S. (2000b) Towards an Ethics for Epersons, in J. Barnden (Ed.) *Proceedings of AISB 2000 Symposium on Artificial Intelligence, Ethics and (Quasi-) Human Rights*, University of Birmingham, 47-52

Torrance, S. (2005). Thin Phenomenality and 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.*

Trevarthen, C. (1994). The self born in intersubjectivity: The psychology of an infant communicating. In U. Neisser (Ed.), *The Perceived Self* (pp. 121-173.): Cambridge Univ. Press.

Varela, F. (1996) Neurophenomenology: A methodological remedy for the Hard Problem. *Journal of Consciousness Studies.* 3 (4). 330-349.

Varela, F. & Shear, J. (Eds.) (1999) *The View from Within: First-person approaches to the study of consciousness.* Thorverton: Imprint Academic.

Varela, F., Thompson, E. & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press, 1991.

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

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

Wegner, D. W. (2002). *The Illusion of Conscious Will*. Cambridge MA: MIT Press.

Zlatev, J. (1999). The Epigenesis of Meaning in Human Beings, and Possibly in Robots*. Lund University Cognitive Studies*, 79.

The Binding Problem: Induction, Integration and Imagination

Susan Stuart*
*Humanities Advanced Technology and Information Institute
University of Glasgow, UK
S.Stuart@philosophy.arts.gla.ac.uk

Abstract

My concern in this paper is with the binding problem and how information, that is stored across the brain, is integrated into one unitary conscious experience, in an act of, what we might refer to as meta-binding. I will draw together the common themes from a diverse body of work that addresses this problem; this work will include Cotterill's neurophysiological approach (Cotterill, 1995, 1998), Kantian metaphysics (Kant, 1929), Sloman's cognitive architecture theory (Sloman, 2004, 2005), Aleksander and Dunmall's engineering approach that entails the integration of cognitive faculties into architectures (Aleksander and Dunmall, 2003), and robotics (Brooks, 1991; Bowling et al., 2004). Fundamental to each of these approaches are the notions of embodiment, animation, perception, and imagination, but, in turn, each of these notions requires a system that has (i) the ability to bind its experiences as experience for it, (ii) the ability to order / tag its experience temporally if it is to be able to plan ahead and direct its attention in an effort to sustain its existence, and (iii) some element of affective processing that makes some things more desirable than others and provides the system with a will to act.

Introduction

There can be little doubt that there is something that it is like to have a perspective on the world, to be a particular human being, weasel or sheep, that isn't simply its being that thing *qua* whatever it is, but which consists, at least in part, in its self-directed interaction with other things in its world. To have a perspective or point of view the system must be a perceiving and, even minimally, conceiving one that is embodied and embedded in its world (Dobbyn and Stuart, 2003). But perceptual and conceptual ability by itself is insufficient for thought, for thoughts need to be conjoined; they must be unified.

When we think of unification with regard to human being we think, though not necessarily in a Cartesian way, of the problematic unity of a body and a mind. The most we can say about ourselves, according to a strict reading of Kant, is that we are logical subjects of thoughts, that we are transcendental unities of apperception that are logically necessary for the very possibility of coherent cognition.[1] We look for the self that draws the perceptual data under some form of conceptual organisation, but find nothing substantial that is the bearer of properties. In its absence we respond by conjuring up a unified self in the concept of a soul or mental thing (Descartes, 1968), in a bundle of discrete perceptions in a mental theatre (Hume, 1739), or we divert attention away from it by proposing a phenomenal unity at a much more fundamental level (Tye, 2003).

Kant was content with neither Descartes' response [*viz.* The Paralogisms of Pure Reason, (Kant, 1929, A339/B397-A405/B432) nor Hume's [*viz.* The Refutation of Idealism, (Kant, 1929, B275-94), and he would be unable to accept Tye's position because it emphasises phenomenal unity at the cost of subject unity, and Kant's position would surely be that for phenomenal unity to take place, that is, to be located, subject unity must first be presupposed.

At first glance Kant's notion of the unified subject as simply a logical subject of thought, the vehicle of concepts (Kant, 1929, B399 /A341), may not seem to be offering us anything more than we have already with Hume. But with a little more care we discover that because of, amongst other things, his requirement that transcendental idealism and empirical realism are logically interdependent, *"The mere, but empirically determined, consciousness of my own existence proves the existence of objects in space outside me.*" (Kant, 1929, B276) and his assertion that *"I am [necessarily] conscious of my own existence as determined in time"* [Ibid.], he must be committed to an embodied and embedded self that is able to perceive itself as enduring through change.

Cotterill (1995) claims that consciousness is primarily associated with movement and response, with the necessary coordination of movement and response requiring a unity of conscious experience. In muscular movement we ask questions about our world, and in its absence – which we can see in people who have lost their proprioceptive sense – we rely on other forms of sensory feedback (Meijsing, 2000). Cotterill suggests that a master node draws together afferent/efferent information into coherent thought and action, and identifies the anterior cingulate as the possible 'site' of consciousness where this activity might come together.

[1] See, for example,

> That the 'I' of apperception, and therefore the 'I' in every act of thought, is one, and cannot be resolved into a plurality of subjects, and consequently signifies a logically simple subject, is something already contained in the very concept of thought, and is therefore an analytic proposition. (Kant, 1929, B407)

Kant's critical philosophy focuses on describing the logically necessary prerequirements for a unity of consciousness, emphasising the role of the cognitive imagination in the act of synthesis. I argue that, like Cotterill, Kant is committed to an active, sensorimotorily enmeshed view of consciousness, so that it must be possible to realise the act of synthesis or binding in some physical system. By examining, albeit briefly, Sloman's cognitive architecture theory (Sloman, 2004, 2005), Aleksander & Dunmall's engineering approach that entails the integration of cognitive faculties into architectures [2003], and the robotics work of Brooks (1991) and Bowling et al. (2004), we can see how this is being done.

Kant's basic architecture of the mind

Kant's account of the mind has two fundamental elements:

1. perceptual awareness or *intuitions* of our world; they are (re)presentations of an object, material passively received by the mind through ones sense organs

2. conceptualisation of perceptual data through the active application of *categories* or *concepts* by the Faculty of Understanding

We can see exactly what Kant means when he claims that these elements are essential for experience by looking at two of his most famous phrases: 'Thoughts without content are empty, intuitions without concepts are blind' (Kant, 1929, A52/B76). If our experience has no content, no experiential or perceptual input, our thoughts will be no thoughts at all for they will be empty. If we have perceptions or experience without any understanding to guide us in our organisation of the data of that experience, we will be as good as blind, for all we will experience, if experience is what we would want to call it, is chaos. So what makes this experience unchaotic – well, two things: (i) the concepts in our understanding that act in some way to synthesise it, drawing together the unity of self-consciousness and the unity of objective experience, and (ii) the power of the imagination. Let us examine these a little further.

Experience is possible only if it refers to an objective world, that is, if we are embedded in an experientially rich and changing environment, for how else would our thoughts have content. And, now it is possible to connect the following claims: 'Unity of diverse experiences in a single consciousness [a self] requires experience of objects' (Strawson, 1999, p.98) and, Kant's argument for the refutation of idealism, 'that there are things in space outside me' (Kant, 1929, B275). To demonstrate this latter claim Kant assumes Descartes' premise that I can determine my empirical consciousness in time without granting the existence of a physical world; but this will fail, for if inner experience is all I could have, then I could never arrive at a conception of myself as a temporally determined consciousness. Thus Kant concludes that inner experience cannot be all there is; there must be an outside world. It is really worth noting here that we see this most clearly set out in the argument for the second Analogy of Experience (Kant, 1929, A189–211/B233–56) where Kant claims that in being able to distinguish surveys from events, that is, in being able to distinguish stasis from movement even when we ourselves are moving, we are aware of our subjective experience as distinguishable from objective fact. It is this capability that makes self-conscious experiences possible.

And now let's look again at the claim that unity of consciousness requires consciousness of unity, that is, to be able to attach the 'I think' to my thoughts they have to be ordered and unified by the application of the concepts and synthesised or brought together by the power of the imagination. This requirement is a bi-directional logical requirement, an interdependence claim, not a contingent relation, and the nomological force of this claim clears the path for claiming that – because we have sensory awareness, understanding, a cognitive or productive imagination[2], and a transcendental unity of apperception it is possible to recognise our thoughts as our own, and all of this is made possible only because there is an external world with which we must engage if we are to have, even an illusory, sense of a continuing self. (Hume, 1739; Brook, 1994; Strawson, 1997, 1999) There is a strong sense in which it is possible to accept that Kant is providing a notion of sensorimotorily enmeshed agent that interacts with its, necessarily changing, world.

Thus, in ordinary cognitive judgement, the manifold of intuitions is 'synthesised' which involves it being brought under concepts to produce judgements. Synthesis occurs through the activity of the productive or cognitive imagination which has three modes: apprehension, reproduction, and recognition. The synthesis of apprehension is *"of representations as modifications of the mind in intuition"*; the synthesis of reproduction is the *"merely empirical law, that representations which have often followed or accompanied one another finally become associated"*, permitting the performance of inductive reasoning; and the synthesis of recognition which is the *"conscious that what we think is the same as what we thought a moment before . . . [without which] . . . all reproduction in the series of representations would be useless. For it would in its present state be a new representation"* and we would be where we were for Hume, the subject of discrete, synchronic experiences. (Kant, 1929, A98-106) Hence ordinary cognition is a product of interaction between the senses, the understanding, and the imagination; it is the conceptualisation and unification of experience with the potential – but only the potential – to be expressed in the form of a judgement beginning: 'I think'. In unpicking this Kantian picture we can begin to clarify what it is that must be resolved if we are to understand the binding problem or how this synthesis operates.

[2] I use productive or cognitive imagination here in opposition to creative imagination, and claim it to be a faculty but rather a power to synthesise or, rather differently, to bring to mind something which is not wholly present.

The Binding Problem

I will begin by setting the problem out in broad terms: knowledge about the world is stored all over the brain. How this knowledge is integrated into one unitary perception to give us conscious experience is called the binding problem. Currently there are two approaches to the resolution of this problem:

1. (i) Space-based binding which claims that there is a specific location or locations in the brain where information is brought together.

2. (ii) Time-based binding which claims that there is no one place where binding happens, because integration occurs over the entire brain and is regulated by some time-based process. Thus, time-based binding looks for when rather than where the binding occurs.

One definition of the binding problem, which is not nearly helpful enough, is given by Valerie Hardcastle (on-line). She says *"Binding refers to the joining together of the individually processed features at the 'psychological' level"* but fails to explain further what is meant by 'the psychological level' and whether, for example, it is conscious or unconscious, or whether it refers to mind states or brain states. I believe Cotterill (1995) overcomes this problem with his use of the term 'plenisentience' and I will develop this in the next section where I will also argue for a hybrid model of temporal and spatial-based binding.

The Role of Attention and Synchrony in Binding

There has been much talk of central executives (Baddeley, 1986) and supervisory attentional systems (SAS) (Norman and Shallice, 1980; Shallice, 1982) and their role in marshalling perceptual input and cognitive processes into making a unified experiential sense of the world; and a great deal has been said about the components that are being marshalled, for example, the visuo-spatial sketchpad and the phonological or articulatory loop; but very little has been said about the particular mechanism that must underpin the functioning of such executive or supervisory systems. In contrast to this Cotterill (1995, 1998) goes out of his way to specify, in some detail, the role of the *master node* in drawing together efferent / afferent information into coherent action and thought. Crucial to Cotterill's theory are both synchrony and attentional processing. It is these we will concentrate on in analysing his theory.

Much of the sensory system works as an outer sense, enabling the organism to determine its external state, and links – directly or indirectly – to actuators, making action, and hence interaction with the world, possible. But in more complex organisms 'sensing' also comprises an 'inner sense', not only enabling the organism to determine its goal(s) and compare its sensory input with its internal state(s), but also to monitor its position, movement and actions in the world. This is the view, a sensorimotorily enmeshed view of conscious expeirence, to which I am committing Kant; it is also Cotterill's view. It is the unity of experience which coordinates information from the senses, including the proprioceptive sense – the ability to sense the position, location, orientation and movement of the body and its parts – with the agent's movement and appropriate responses.

In both Kant and Cotterill we see an emphasis on attention to movement for it is in attending to changes in our world, including changes in our body and position, being in a state of – what Cotterill describes as – 'plenisentience', where inputs can be consciously sensed and unconsciously processed, that with, for example, the proper functioning of cells in the visual cortex, we are able to distinguish movement from stasis.[3] But Cotterill adds to this that the position of our muscles and our subsequent muscular movements are what makes it possible for us to ask questions about our environment and our position in our environment. [Cotterill 1995, p.297] In fact we can go much further than this, for if we were unable to move or unable to receive feedback through our senses, our interaction and knowledge of our world would be very limited. If you are a tree this matters little but if you are an animal that must avoid being prey and instead become predator it is essential. Thus, if our sensory system including our proprioceptive sense is working well we will be in a position to receive and translate afferent signals and produce appropriate efferent impulses in response. In this way we become aware of our world and, through the development of a body schema, are able to conceive of ourselves in relation to, whilst still being autonomous within, our world.

We have a ready made counter-example to any opposition to this claim in those individuals who have lost their proprioceptive sense, for without their internal feedback system they rely on the external feedback provided by, in most cases, their visual sense to regain their sense of selfhood or identity. Meijsing (2000) says of the patient, Ian Waterman, 'In the dark he did not know where his hand was; and even if he knew, he would not have been able to move it towards the bedside table without visual feedback' [p.42]. Yet, even

[3] In humans and other primates, the vestibulo-ocular reflex (Churchland and Sejnowski, 1992) operates by direct feedback between sensory units (the semicircular canals) and actuators (motor neurons in the eye) with no 'inner' representational or cognitive system intervening. Light in the eye falls on the retina and, depending on its intensity and wavelength or colour, is translated by rod and cone cells into electrical impulses which are then transmitted along the optic nerves to the visual cortex at the back of the brain. It is in the visual cortex that this information is translated into perception of colour, depth, objects and movement. The lens and retina act in some ways like a camera but the information that is transferred onwards to the visual cortex is in a different form altogether. There is no single visual cortex, rather there are assemblies of discrete cells some dedicated to discerning edges, some to motion, some to colour, and so on. Neuropathological evidence shows us that damage to one batch of cells leaves others unaffected. For example, damage to the 'colour cells' will leave the individual able only to perceive in monochrome, and damage to the cells that determine objects might leave the individual able to perceive motion but without objects!

this is insufficient for a fully unified sense of self.[4] Ian Waterman's sense of unity, his coherent sense of self, returned only when he had learned to move again with a great deal of concentrated visual feedback. Thus, it is is not just the passively received information about a changing environment, but the interplay between this information and active self-movement that places the self, a unity of experience, firmly at the centre of its environment. It is active self-movement which gives a sense of agency, as the perceived environment changes as a result of the agent's purposive action (Meijsing, 2000, p46).

Miall and Wolpert (1993) state that the brain structure should

> ...receive as inputs an efferent copy of the motor command being sent to the ...limb, and also proprioceptive information about the current state of the body. The latter is needed for an accurate internal representation of the limb, as the arm's mechanical properties depend on its position and motion. Hence, the internal dynamic model must be updated by proprioceptors. [p.209]

If it isn't, then, like Ian Waterman, we lose our means of unifying our experience, and over – a fairly short – time, we lose our sense of self.

Cotterill suggests that there is a hierarchy of muscular control over which there is a global control mechanism – which he justifies on the grounds that there are limits to the amount of information that can be handled by the system at any one time (Broadbent, 1958) – and, in accordance with the spatial-based binding approach, he proposes that it be the anterior cingulate that acts as this global control mechanism because it is neurally close to the higher motor hierarchical levels in the brain and because it is here – according to evidence from positron emission tomography [PET] scans [See Pardo et al. (1990) – that a response is translated into a physical or motor directive. For example,

> When I recognize a lemon, I am simultaneously detecting its pointed-oval shape, its dimpled skin, its yellow colour, and possibly also its relative softness and its characteristic citrus smell. The first three of these attributes are all detected by the visual system, but by different parts of it. The fourth feature is detected by touch, while the last one is discerned by my olfactory sense. And where is the logical conclusion, *lemon*, located? It is nor deposited in some inner sanctum, farther up the hierarchy. On the contrary, its components are left in those same sensory modalities and areas. The concept *lemon* merely exists through the temporary binding together of its various attributes; and we are able to sense the lemon as a single unity because we can instantaneously detect what goes with what. (Cotterill, 1995, p305)

Kant would complete this last sentence by saying '...and we are able to sense the lemon as a single unity because we draw together the intuitions under concepts and with the synthesising power of the apprehensive, reproductive, and recognitional imagination we are able to put together a thought which might be 'lemon' or, more complicatedly, 'I think it is a lemon'[5], and being able to unite the disparate parts of my perceptual experience together into a thought is sufficient to reveal that the thought is being had in one head, that is, that it is 'my thought'.

In his emphasis on the synchrony of input and output Cotterill presents us with a temporal-based response to the binding problem, but only in some circumstances. In the process of object recognition there is a great deal of neural activity which is a result of the information received through the modalities involved in the perception of an object, and that neural activity is distributed across the parietal-temporal-occipital association cortex. But we have also seen that synchrony of input alone cannot be the complete or sole explanation for binding. The detection of movement, both internal to the system and external to it, and the co-ordination of sensory input, including the proprioceptive sense, with the agent's movement and responses, are essential if there is to be a unity of consciousness, and a unity of consciousness is necessary if we are to have a coherent experience of our world. If Cotterill is right, this aspect of the neural activity must be located at the centre for determining muscular direction in the brain, for it is with the proper functioning of the anterior cingulate, that we can ask questions about, and bring about changes within, our dynamically changing environment. Cotterill's picture is, then, of a hybrid model of how binding occurs; it is time-based in its synchrony and it is space-based in that it might be located in the anterior cingulate.

Cognitive Architecture Theories and Robotics

Early hybrid cognitive architectures represented knowledge symbolically as rules and facts but had a neurally-based activation process that determined which facts and rules got deployed in which situations. [See ACT-R and SOAR, (Anderson, 1983, 1990, 1993).] Sloman's Cog-Aff and H Cog-Aff architectures provide a more holistic approach to the requirements for consciousness experience, arguing that structural, cognitive and affective components must be combined in one architecture – one subject of consciousness – consisting of physiological machine sub-architectures and virtual sub-architectures of mental states and mechanisms. Sloman distinguishes between three types of processing: perception,

[4] She goes on to note that this is the replacement of an inner sense with an outer sense, and these may not equate to the same thing: the inner sense seems to be immune to error through misidentification [cf. (Evans, 1982), (Brewer, 1995).

[5] But the more complicated utterance 'I think it is a lemon' would surely only be uttered where there is some doubt about the object's status.

action, and central processing, each recognisable in Kant and Cotterill, and three cognitive levels: reaction, deliberation and reflection. Reaction, he argues, and few would disagree, is the oldest part of any cognitive architecture, with deliberative or inferential reasoning coming later, and finally, the 'meta-management' of reflection emerging much later still and, possibly, only in human conscious agents.

Unlike early architectures, H Cog-Aff is not algorithm- and representation-based, which is bound to be a distinct advantage when developing virtual architectures, and even though Sloman raises numerous important questions about the nature and ontology of meta-management constraints like the emotions, he offers no account of the central processing element that acts to bind the perceptual information with other cognitive mechanisms and affective attitudes. But criticism of this sort might be too harsh, for it is clear that the complexity of the cognitive system does not lend itself to easy explanations. As Sloman says, it isn't *"a single (atomic) state which switches when some input is received . . . There may still be real, causally efficacious, internal virtual machine events and processes that cannot be directly observed and whose effects may not even be indirectly manifested externally"*(Sloman, 2005). In response he suggests that we might think about virtual architectures in terms of *"multiple concurrently active, interactive, sub-states changing on different time scales (some continuously) with varying complexity"* [Ibid.], and if he is right, then it will be some time before Cog-Aff and H Cog-Aff architectures have anything to say about their central processor.

Aleksander and Dunmall (2003) present, in axiomatic form, a formal statement of five mechanisms that are thought minimally necessary to underpin consciousness and, thus, the creation of conscious machines. Theirs is an engineering approach that entails the integration of cognitive faculties – perception, imagination, attention, prediction, and emotion – into computer-based depictions to create sensations of 'outthere'. That these elements appear here as they have in Sloman and, to some extent, Kant and Cotterill is no surprise, but what is particularly novel about Aleksander & Dunmall's approach from the point of view of this paper is that it presents a sort of Kantian transcendental argument against the zombie theorists who argue that qualia or sensation cannot be supervenient on mechanism. Aleksander & Dunmall turn the argument on itself and show that mechanism is implied by the occurrence of sensation. In short, mechanism may not imply sensation, but sensation implies mechanism. That we would find an argument of this sort in Aleksander's work is not surprising since he asks the same question Kant asked but from an implementational point of view: What are the essential mechanisms for being conscious? In another Kantian twist he concludes that the emergence of self results from a combination of sensory, imaginational, attentional and (though Kant might be forgiven for excluding this element) affective depictions which can start with the logical subject 'I'. Neither Cotterill nor, I feel, Sloman would object.

Brooks' work on cognitive architectures and robotics (Brooks, 1991) is rather flat, dealing with perceptual and reactive processes only.

> . . . (the robots) are situated in the world – they do not deal with abstract descriptions but with the here and now of the world directly influencing the behaviour of the system. . . (Brooks, 1991, p575)

Perception and action are connected directly; there is no central processing system, no central representation, and whatever binding there might be said to be must be temporally-based. Where Sloman argues for a combination of physical and virtual architectures, and a move away from the image of a single state mechanism, Brooks' animat architectures have been strictly engineered: the finite state machines that govern their low-level behaviour have been carefully contrived; and the patterns of connection and message passing between these machines are the result of much experiment. These are behaviour-based machines and there is little thought for affective processing, or the role played by the imagination.

In contrast to Brooks' work is Bowling, Browning, & Veloso's robotics work (Bowling et al., 2004). Their concern is with the use of neural networks to produce the kind of unpredictable behaviour involved in the dynamic environment of robot soccer; as they say:

> [The] challenge of controlling a team of robots within the context of robot soccer, a multi-robot, goal-driven, adversarial domain.

In this complex environment the binding of experiential input must occur within each individual as well as across the team if they are to achieve their end and score goals or, at the very least, block their opponents goal scoring. So,

> Given a set of effective and parameterized individual robot behaviors, how do we select each robots behavior (possibly using past execution experience) to achieve the teams goals?

The elements necessary for successful play are perception, attention, reaction, coordination, and prediction, all of which are made possible by the fusing or integration – the binding – of information which operates on the basis of probability algorithms. Such algorithms must occur in both a temporal and a spatial framework, the *playbook* which is "a method for seamlessly combining multiple team plans", if the robots are to act appropriately and effectively in real time. Thus, the Bowling et al. (2004) soccer robots[6] operate on a hybrid model of temporal and spatial binding.

[6]The Bowling et al. (2004) soccer robots – the most sophiticated being the legged-teams using Sony AIBOs – have a variety of names from CMPack to CMDragons; information about them can be found at `http://www-2.cs.cmu.edu/~robosoccer/main/`

Concluding Remarks

There are many common elements within these seemingly disparate approaches, with the exceptional case being Brooks. The others have at their core the notions of embodiment, animation, perception, and imagination, and for their instantiation they require a system that has the ability to synthesise its experiences and be able to recognise them as experiences for it. Without this there can be no urge to act, for unless I am aware of experiences being mine – whether rat, cat, badger, or soccer robot – I will have no desire to act to defend myself from predators or to act stealthily towards prey. For this reason I must be able to tag my experience temporally, not only to enable me to recognise that especially vivacious experiences are current and require immediate action, but also to make possible associationistic learning and the construction of preference models of those things which are desirable and those things which are not.

Cotterill's account of consciousness is primarily associated with movement and response, and the co-ordination of movement and response requires a unity of consciousness or subject unity. This is an interdependence claim not unlike Kant's claim that a unity of consciousness is possible only on condition that we have a consciousness of unity, and *vice versa*, and we have seen that Kant's argument for our being able to conceive of ourselves as unities of consciousness, that is, as temporally determined conscious agents, is based on our being able to discern and distinguish movement from stasis through the application of *a priori* concepts which order and unify our perceptual input. A great deal is implicit in Kant's notion of ordering and unifying, a great deal that is being excavated by current work in neurophysiology, robotics and cognitive architecture theories.

Cotterill's argument focuses on a neurophysiological approach to the problem and identifies the anterior cingulate as a possible 'site' of consciousness. Sloman's cognitive architecture theory is a model inspired by work in artificial intelligence. Aleksander & Dunmall's approach is combination of mathematics and engineering; and Bowling, Browning & Veloso's approach is based in Artificial Neural Networks (ANN). None of these approaches were available to Kant, yet in his metaphysical enquiry we find him committed to an active, sensorimotorily enmeshed view of consciousness, a view which is not just recognisable in, but there as an underpinning to, each of the very different frameworks of enquiry addressing the problems of consciousness and the integration of thought.

References

Igor Aleksander and B Dunmall. Axioms and tests for the presence of minimal consciousness in agents. *Journal of Consciousness Studies*, 10(4-5), 2003.

J R Anderson. *The Architecture of Cognition*. Harvard University Press, Cambridge, MA, 1983.

J R Anderson. *The Adaptive Character of Thought*. Erlbaum, Hillsdale, NJ, 1990.

J R Anderson. *Rules of the Mind*. Erlbaum, Hillsdale, NJ, 1993.

A D Baddeley. *Working Memory*. Clarendon Press, Oxford, 1986.

M Bowling, B Browning, and M Veloso. Plays as effective multiagent plans enabling opponent-adaptive play selection. In *Proceedings of International Conference on Automated Planning and Scheduling (ICAPS'04)*, 2004.

B Brewer. *The Body and the Self*, chapter Bodily Awareness and the Self. MIT Press, Cambridge, MA, 1995.

D E Broadbent. *Perception and Communication*. Oxford University Press, Oxford, 1958.

A Brook. *Kant and the Mind*. Cambridge University Press, 1994.

Rodney Brooks. Intelligence without representation. *Artificial Intelligence*, 47(1-3):139–159, 1991.

Patricia S Churchland and T J Sejnowski. *The Computational Brain*. MIT Press, 1992.

Rodney M J Cotterill. On the unity of conscious experience. *Journal of Consciousness Studies*, 2(4-5):290–311, 1995.

Rodney M J Cotterill. *Enchanted Looms: Conscious Networks in Brains and Computers*. Cambridge University Press, 1998.

René Descartes. *Discourse on Method, and the Meditations*. Penguin, 1968.

Chris Dobbyn and Susan A J Stuart. The self as an embedded agent. *Minds and Machines*, 13(2):187–201, 2003.

G Evans. *The Varieties of Reference*. Clarendon Press, Oxford, 1982.

David Hume. *A Treatise of Human Nature*. Clarendon Press, Oxford, 1739.

Immanuel Kant. *The Critique of Pure Reason*. Macmillan Press, 1929.

M Meijsing. Self-consciousness and the body. *Journal of Consciousness Studies*, 7(6):34–52, 2000.

R C Miall and D J Wolpert. Is the cerebellum a smith predictor? *Journal of Motor Behaviour*, 25(3):203–216, 1993.

D A Norman and T Shallice. Attention to action: Willed and cutomatic control of behavior. CHIP Report 99, University of California, San Diego, 1980.

J V Pardo, P J Pardo, K W Janer, and M E Raichle. The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm. *Proceedings of the National Academy of Sciences*, 87:256–9, 1990.

T Shallice. Specific impairments of planning. *Philosophical Transactions of the Royal Society of London*, B298:199–209, 1982.

Aaron Sloman. Varieties of affect and learning in a complete human-like architecture. online, July 2004. URL `http://www.cs.bham.ac.uk/research/cogaff/talks/#talk24`.

Aaron Sloman. What are information-processing machines? what are information-processing vitual machines. online, January 2005. URL `http://www.cs.bham.ac.uk/~axs/misc/talks/information.pdf`.

Galen Strawson. The self. *Journal of Consciousness Studies*, 4(5/6):405–428, 1997.

Galen Strawson. The self and the SESMET. *Journal of Consciousness Studies*, 6(4):99–135, 1999.

M Tye. *Consciousness and Persons: Unity and Identity*. MIT Press, Cambridge, MA, 2003.

You Only Live Twice; Imagination in Conscious Machines

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

Abstract

The role of imagination in perception, memory and consciousness is considered. Imagination is seen as an essential part in perception, cognition and memories. The view of a conscious machine as a perception driven system is rejected. Instead it is proposed that a conscious machine should have an imagination-augmented inner life that actively seeks to perceive the world according to its own needs. The meanings of the imagined entities must be grounded to real world entities. This can be done easily by using the perception circuits also for imagination. This kind of re-entry will also facilitate the continuous flow of inner speech and imagery. The enabling functions of imagination as well as the higher-level structure of thought and imagination and their relation to consciousness are considered and requirements for the supporting mechanisms and system architecture are outlined.

1. Introduction

A machine, in order to be conscious, has to be able to perceive the world, its bodily self and also the flow of its own mental content. The flow of the mental content, the machine thoughts, can be easily assumed to arise as a response to the flow of sensory percepts. However, this view may be too limited.

Our thoughts are not necessarily related to the concurrent sensory percepts. Instead, they may be about completely different, imagined matters, only to be disturbed and interrupted by strong sensory stimuli. This inner life is an essential part of the contents of consciousness. This has been recognized for instance by Hesslow. He sees thinking as the mental simulation of actions and perception with anticipation. This, according to Hesslow, leads to a conscious inner world that does not immediately depend on external input. (Hesslow 2001, 2002).

Thus, the cognitive system is faced with two aspects; the external world appearing as the flow of sensory percepts and the inner life that manifests itself as the flow of inner speech, imagery and "the feel of the moment".

At each moment we feel something. We feel warm, cold, bodily comfort or discomfort, tired, hunger, thirst, etc. These form the most fundamental part of our inner life and guide our perception of the external world; what should be done in order to achieve a comfortable state.

Therefore, in our quest for conscious machines, should we first seek to create a machine with inner life or should we seek to create a perceptive machine with motor skills and responses and see if any inner life would arise? The answer is not, naturally, yes or no. When designing a conscious machine we must simultaneously consider all sides of the process.

It is a straightforward task to design a robot that reacts to sensory stimuli, is even able to learn something and can adjust its responses accordingly. It is another thing to design a robot that also generates an inner life and actively seeks to perceive the external world according to the needs of this inner life. Yet it is this inner life that would provide a robot with a "self", "mind" and personality.

Traditional robots have very little in the way of imagination and inner life. Recently though there have been some attempts towards that direction. Holland has investigated robots with internal models (Holland and Goodman 2003). Nevertheless, robots with true inner life remain yet to be demonstrated.

It is obvious that imagination is a content-level phenomenon and as such may be described without explicit references to the hardware that supports it. Yet, even though we all are able to imagine things the actual processes of imagination are not clear, apart from some trivial cases. How could we then determine the requirements for imagination-supporting circuitry in a machine? What kind of signals and signal sequences should the system gen-

erate if it were to perceive imagined objects and actions? What kind of memory and cross-connection architecture would allow imagination representations to behave in a similar way to those representations that are caused by perceived real world events? What kind of events would trigger and initiate imagination and what kind of processes and criteria would determine the course of imagination? What would be the mechanism for the moment-to-moment connection between perception and imagination? How could the machine make the difference between the "make believe", the "real but not present" and the "present real"? How could "the feel of the moment" be created in a machine?

2. About Imagination

Imagination is the forming and manipulation of mental representations of actions and entities, which are not sensorily present. However, pure memory recall, similar to tape recorder playback, would not be considered as imagination here.

Imagination and perception are connected. Our senses produce only a limited amount of attended information. We do not see behind our backs, we do not see behind objects, yet we can imagine what is there. We can hear sounds and we can imagine the source of these. We can see what a person does, but we will have to imagine the reason for these acts.

It has been proposed and there seems to be some experimental proof that we learn to make sense of percepts via explorative actions (Gregory 2004, O'Regan and Noë 2001, Taylor 1999, pp. 212 – 218). This would later on enable us to imagine these actions and the information that might be revealed if the explorative actions were actually executed. Thus, for instance, the retinal images of objects do not necessarily carry much of information about the actual objects, instead they may be more like fuzzy symbols that evoke the related information that we already have associated with the same. One might even say that objects are not really recognized; they only remind us of something; something that they might be, something that they might allow us to do.

Thus, we do not perceive the world as an odd collection of entities, instead we perceive possibilities for action that these entities suggest to us. Gibson (1966) introduced the concept of "affordances" to describe that very situation. Gibson seemed to understand that the external world would offer these possibilities directly, without any top-down cognitive processing. However, stones do not broadcast: "Use me as a hammer". In reality the perceived affordances are products of association and imagination.

Thus perception is not passive reception of sensory information, instead it is an active process involving exploration, anticipation and imagination. This view is also expressed in the perceptual activity theory (Thomas 1999).

Imagination and memories are related, too. After an event we have only memories of the perceived episode, but already the original percepts were laced with imagination. In fact our whole picture of the external world may be based more on imagination and less on accurate sensory perception and precise memories. Yet this imagined version or expectation about the world usually coincides quite well with the real world when we test it. We imagine what to do next, what to do in the future and we try to adjust our behavior in order to achieve these goals. We can use imagination for creative purposes.

Both recalled and imagined representations originate from inside. Therefore, what would be the actual difference between memories and imagination? Pure memories, similar to tape recorder playback may not exist. Recalled memories are not recordings of sensory percepts but imperfectly reconstructed records of past contents of consciousness, including thus products of imagination.

What is imagination technically? Is it a succession of mental pictures or abstract tokens or what? There are various theories but none of them seem to be completely satisfactory for the time being. A low-level theory would explain the material mechanisms and carrying representations while a higher-level theory would explain how the content of imagination behaves.

How high level should a theory of imagination be? Imagination manifests itself at content-level and therefore there is a built-in booby trap there. A theory that tries to deduct the fundamental carrying mechanisms of imagination by the analysis of the content matter may fail, because the sought-after rules may no longer be visible at that level. It is as if trying to explain the workings of radio by inspecting received programs.

Nevertheless, we need also a higher-level theory of imagination for the construction of conscious machines. Once we have that then we can envision various platforms that can support the designated processes. We can by-pass the philosophical ruminations about mental pictures, tokens and the like, as it would be a rather straightforward engineering task to design suitable representation methods and circuits for these.

Imagination may be divided into two classes: 1.) Imagination of the possible; entities that could exist, actions that could be executed. 2.) Imagination of the impossible, entities that could not exist, non-executable make-believe fantasies.

How does the system know the difference between the possible and fantasy? Is there an exact line between these? I can imagine getting up and having a cup of bad coffee, I can also imagine that

in the afternoon I will be a king. One of these imaginations is pleasant, the other is not. Likewise, one of these imaginations is possible, the other is not, but which kind of mechanism would tell me the credibility of each imagination? Obviously this mechanism would not be based on the pleasantness of the imagination.

Such mechanism might be based on familiarity and experience. If something similar has happened before then it might be realizable and true also in the future. On neural level this would correspond to a match-operation; the imagined entity would be matched against memories. However, on neural level the memories would appear as representations that are not necessarily different from those that have been generated by imagination. Therefore some further criteria would be necessary to make the distinction between memories of true events and memories of imagined events. It would help if there was some kind of true and false, make-believe value that could be associated with memories of true events and memories of imaginations.

Imagination and the distinction between the real and the make-believe are also related to the symbolic use of tokens. In children's play toys may not necessarily bear any resemblance to the objects that they are taken to depict. Instead, they are only used as tokens that help to mark the relationships between the entities of the play and communicate these to the playmates. The actual play takes place inside the head! A stone is not a car and the child knows that, but for the play the stone may temporarily carry associations that would be there if the stone was indeed a car. Thus percepts and inner representations are not only taken to represent the corresponding actual and direct external entities but may also be used to stand for something completely different.

3. Imagination and Consciousness in Machines

Imagination and consciousness are related. Imagination does not only involve the creation of mental representations, it also involves the bringing of these imagined representations into our awareness. Thus the contents of our consciousness contain products of imagination. Still, there is a deeper connection between imagination and consciousness. We can imagine not only external entities that might be sensorily perceived, but also our own motor acts and behavior. In doing that we must be able to imagine ourselves doing something and this, in turn, necessitates some kind of mental self-image and self-consciousness. Therefore, without imagination the content and scope of consciousness would be severely limited.

Thus, a supposedly conscious machine should also possess the capacity of imagination. This has already been realized by, for instance, Aleksander & Dunmall (2003). They have included the requirement of imagination in their list of criteria for conscious machines. According to these criteria a conscious machine shall have internal states that depict entities that are not present and moreover, shall have means to control imaginational state sequences to plan actions.

Incidentally the requirements of Aleksander and Dunmall conform to the specifications of the author's description of a cognitive machine. The author has even proposed that if the existence of the flow of mental content in the machine; especially inner speech and inner imagery, could be monitored and thus verified and if the machine itself could report having this content and would recognize it as the product of its own imagination then the machine should be deemed to be conscious. (Haikonen 2003, 2005).

In the following requirements for a cognitive system with imagination are discussed.

4. On the Enabling Functions of Imagination

A cognitive system that is supposed to have the ability of imagination must provide and support a set of prerequisite or enabling functions. The enabling functions of imagination are taken to be here:

1) The evocation of mental representations of imaginary objects at different positions; what, where
2) The evocation of mental representations of change and motion
3) The evocation of mental representations of relation: Relative position, relative size, etc. and relative motion; collide, pass by, take, give, etc.
4) Mental modification: Make larger, smaller, rotate, combine, move from one position to another, etc.
5) Attention, introspection
6) Decision making; the selection between competing imagined scenarios

The basic system requirements for imaginative machines may be deducted from these functions.

The functions 1 – 4 involve the representation of various properties of entities and their modification. It must be especially noted that also self-percepts, such as the position of various body parts and their movements are included here, otherwise the imagination of the same would not be possible.

In principle there are two methods of representation; representations with and without fine structure. Distributed representation is an example of a repre-

sentation method with fine structure (Hinton et al. 1990) while the "grandmother" signal representation and the unique symbol representation would be examples of representations without fine structure. Obviously representations with fine structure would be useful for imagination; tweaking the fine structure would allow the generation of representations of modified entities. On the other hand, "grandmother" signals offer the ultimate compression of the representation, but also necessitate dedicated wiring with growing complexity as the number of entities to be represented grows. Thus it is not feasible to have a distinct symbol or signal for every possible entity and state of affairs. The optimum system economy may be achieved via a mix of representations with and without fine structure.

Representations with fine structure may be constructed as arrays of "grandmother" signals for a large but limited collection of elementary properties or features. These "grandmother" signals could be used as the "set of alphabets" for the representation of the world. Thus each represented entity would be a combination of these "alphabets". It is known that a subset of any set can be larger than the original set. Therefore a mechanism that could create subsets of subsets would allow the representation of unlimited number of entities by a limited number of "alphabets".

What would these "alphabets" be? It seems that perception utilizes its own "alphabets". It is known that auditory perception is based on the frequency analysis of the heard sound, thus the auditory "alphabets" would consist of frequency domain entities and a perceived sound would consist of temporal sequences of these. Likewise, visual perception seems to be based on the detection of elementary visual features. In introspection imaginations are not unlike something that we might pretend to perceive and therefore it would be tempting to claim that the "alphabets" of imagination were the same as the ones of perception. In cognitive machine design we must consider this proposition by the resulting performance and economies of circuitry.

Barsalou (1999) has argued that cognition must be based on perceptual symbols, which are modal and analogical, componential and not discrete. Here the representations consisting of combinations of the "alphabets" would seem to fall into this category. However, it should be noted that via association these representations may be set to represent completely different entities, ones that are not in the least way analogical to the actual representation.

The tentative equation of the "alphabets" of perception and imagination leads to an economically minimized design where the perceptual circuitry is reused by imagination processes. This can be achieved via suitable feedback (reentrant) loops, see for example Haikonen (1999, 2000, 2003). There is some indication that this is the situation also within the human brain. Le Bihan, Kosslyn, Davidson & Schwartz and others seem to have experimental proof that visual imagination indeed utilizes the same brain areas as vision (Hesslow 2001, 2002).

The principle of a feedback loop system that allows the perception of imaginations by the perceptual circuitry is depicted in the figure 1.

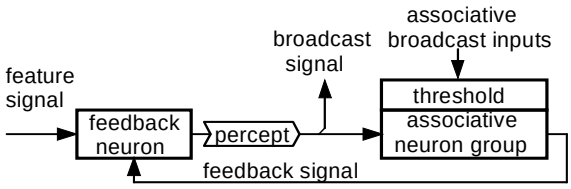


Figure 1. A perceptual feedback loop

In the figure 1 a feature signal that constitutes one of the "alphabets" of perception arrives from a sensor. The presence of the feature signal indicates the presence of the depicted feature. The feature signal as a percept may be activated either by the corresponding sensed property or by the feedback signal from the system. In both cases the basic meaning of the percept is that of the feature signal. The percept signal is broadcast to the rest of the system. The feedback signal may be associatively activated by the associative broadcast input signals. In this kind of loop the percept signal represents imagination whenever it is activated by the feedback and no sensory signal is present. The feedback signal may also be used to prime the sensory signal. Primed sensory signals can be made to have higher intensity and may therefore more easily pass various thresholds in the system. The actual system would consist of a very large number of these perceptual loops laid in a parallel way.

Would it be possible to imagine things without actually having to circulate large arrays of feature signals (the alphabets)? This may be so; imagine grasping a familiar object with your hand, eyes closed. The changes are that you are able to open your fingers just correctly without any clear mental image of the object. Thus it would seem that the full array of feature signals are not used, yet precision motor acts can be initiated. Indeed, circuitwise it would be more economical to manipulate smaller signal arrays instead of large arrays of feature signals. On the other hand the system must be able to evoke a large array of effector ("finger position") signals in order to grasp accurately the imagined object. This can be achieved by associating a smaller "token" signal array with the actual feature signal array and using this token in the imagination process. This leads to an architecture like that of the figure 2.

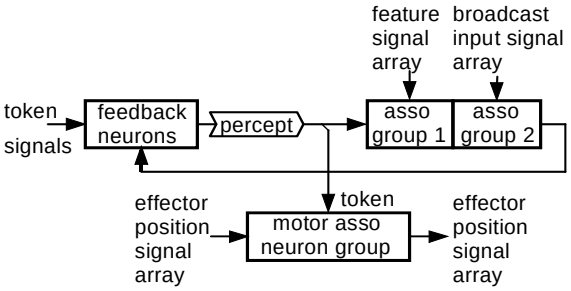


Figure 2. Token signals as the means of imagination

In the figure 2 the token signals may be a small subset of sensory signals. The feature signals that describe the intended object are associated with these token signals. Thereafter the token signals may be used instead of the large feature signal arrays and may, for instance, be evoked by broadcast input signals. The effector position signal array that describes the opening of fingers is also associated with the token signals and may therefore be also evoked by the same, thus the complete feature signal array is not needed and a fuzzy mental image may still lead to precision motor functions. This is also related to the fact that it is easier to recognize objects than describe them.

It is not sufficient only to be able to evoke representations of imagined entities at various sensory input locations. Therefore the item 5 of the list of enabling functions contains attention and introspection. The system must also be able to inspect the evoked imagery or linguistic construct. This calls for temporary storage of the evoked representations and the possibility of attentive selection. This requirement is illustrated by the figure 3.

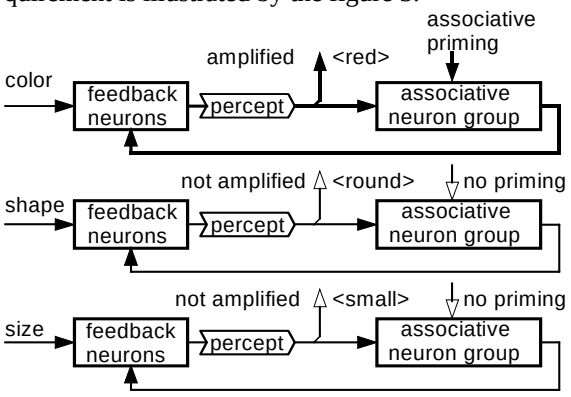


Figure 3. Attention; selection via priming

In the figure 3 perceptual feedback loops for the visual properties of color, shape and size are depicted. In this case each of these loops would actually consist of a large number of single signal loops of the type of the figure 1.

This system must have the means for attentive inspection of imagined entities. For instance, the machine may be asked: What is the color of a tomato? As the response the machine must be able to evoke at least a vague mental representation of a tomato. In terms of distributed representations a tomato would be an entity with the properties <small>, <round> and <red>. These would be represented by separate signals at the separate locations for size, shape, color, etc. The system should be able to sustain these signals as long as needed. Now the quality of one property, color, is requested; the corresponding property signal should stand out among the rest. This can be achieved via priming. The word "color" would prime the color loop and thus the color percept <red> would be broadcast at a higher level. In this way attention would be brought on the correct answer.

The function 6, decision-making involves the evocation alternatives and the subsequent selection between these. A situation may evoke an imagined response, which may be executed. Here it must be decided whether the response is to be executed or not. Another situation may evoke several possible responses. Now it must be decided which one of these responses is to be selected. It is obvious that some kinds of decision criteria are needed. Should I go to the movies or stay at home and watch TV? Without any further arguments these two scenarios might be of equal value and the decision could only be a random one. However, we may have additional points to consider. The movie is good and TV is dull, on the other hand it is raining and I am rather tired... These arguments are not neutral; they carry emotional value, which can be used as the basis for the decision.

Thus, imagination must be combined with an emotional value system. This system must evaluate the good/bad, pleasant/unpleasant value of the imaginations and use this as an attention or selection mechanism. A possible artificial emotional value system mechanism has been proposed by the author (Haikonen 2003). According to this proposal good/bad and pleasant/unpleasant values are grounded to specific sensory inputs and these sensations have specific hard-wired system reactions, which also control attention. These "emotional" values may also be associated to other percepts and memorized along these. Therefore also recalled or imagined representations bring forth their associated emotional values, which then will affect attention and decision making.

5. On The Structure of Thought and Imagination

In the previous chapter the inner representation and manipulation of entities were considered. However, human thinking and imagination involves

more than the representation of isolated entities. A thought is more than that. It may have the shape of a description, comment, question, or a command. Here inner speech is considered as a form of thought typical to humans. Steels (2003) has emphasized the cognitive importance of the inner speech and has proposed that inner speech was a side effect of being able to learn and use external language. This is also the author's view (Haikonen 2003).

The strange thing about the inner speech is that the listener and the speaker are the same person. Usually the speaker knows what he is going to say; if the listener is the same as the speaker, then also the listener should know. Therefore, why to say anything at all?

This paradox will dissolve if we assume that instead of a "conscious speaker" there is only a subconscious process that generates thoughts that may only be consciously perceived if they are transformed into the equivalent of heard speech. (Obviously the same would apply to visual imagery.)

It is proposed here that a present thought is the result of a subconscious process that evokes a number of preparatory candidate thoughts. This process is depicted in the figure 4.

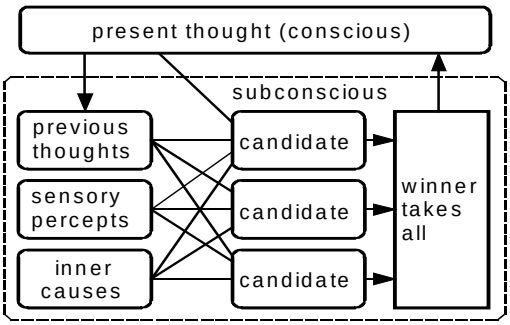


Figure 4. The subconscious generation of thoughts

The candidate thoughts are associatively evoked by the present and previous thoughts, the concurrent constellation of sensory percepts and/or inner causes; such as bodily needs, emotional states, etc. The most intense candidate is supposed to become the next present thought via a "winner-takes-all" threshold process. The present thought is conscious and as such can be reported; in fact it is the report. As such it is broadcast to the entire system and is therefore able to affect the system as a whole, while the candidate thoughts are not conscious and cannot be reported and are not globally broadcast.

Sensory percepts are evaluated for their significance and emotional value. Important percepts will gain full attention, good and bad values will lead to accept/reject and approach/withdraw reactions. Pleasant/unpleasant percepts will lead to continuation/discontinuation of the related action. The reuse of sensory perception circuits for the perception of thoughts makes these evaluation and reaction processes available also for the internally generated thoughts.

This process would seem to be a completely automatic one. However, here we find the interface between the content and carrying processes, between the conscious and the subconscious. The content affects the outcome of the process of figure 4. To illustrate this let's consider a possible chain of thought: "If I do like this then... no, no... it wont work... I have to think something else... what.. what.. maybe this way.." We are aware of the results of the match/mismatch and emotional evaluation processes and are therefore able to perceive the need to generate new thoughts that would better correspond to our mental goals. This evaluation would prime the subconscious processes that evoke candidate thoughts.

The resulting chain of conscious thoughts is not necessarily a linear one, with a new thought following another. A chain of thought may not lead to a satisfactory mental outcome, therefore it should be possible to return to the starting point and initiate another line of thought as depicted in the figure 5.

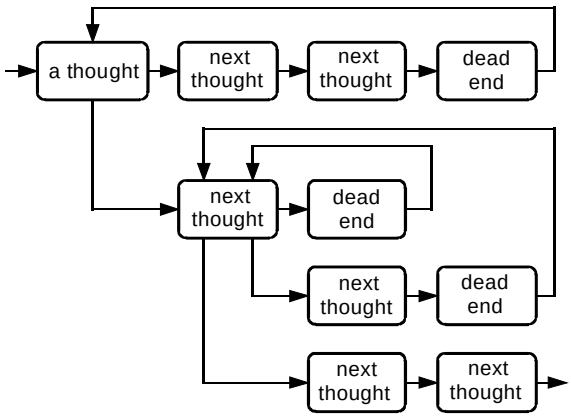


Figure 5. The conscious generation of the flow of thoughts

This kind of looped thought process has implications for the machine architecture. Short-term memories are needed for the possible returning to the starting thought so that a new chain of thoughts could be initiated. Additionally there should be a mechanism that would suppress previously selected candidates so that reruns of previous thought chains could be prevented.

6. Conclusions

A conscious machine has to be able to perceive the world, its bodily self and also the flow of its own mental content. Perception is not a mere passive

reception of sensory signals, instead it is an active process that seeks and interprets information according to the needs of the cognitive system. The world is seen as affordances, which are created by imagination. A supposedly conscious machine should have an inner life, the flow of meaningful mental content. The capacity of imagination calls for inner representations that can be evoked by sensory information and by inner causes only. These representations must have fine structure that allows the modification of the appearance of the imagined entities. Feedback (re-entrant) architectures allow the re-use of sensory perception circuits for the perception of imaginations, too.

The serial flow of inner speech and imagery are seen as running reports of the results of subconscious and parallel processes that seek suitable responses for the instantaneous situation. The interface between the conscious and the subconscious is seen here. The direction of the thought flow can be consciously altered; the new thoughts arise however from subconscious processes.

You only live twice, once for your inner life and once for the external reality. Imagination is the key to consciousness and inner life. Conscious machines should have an inner life and dreams, too.

Acknowledgements

The related work is done at Nokia Research Center (NRC). The author wishes to thank the head of NRC, Dr. Bob Iannucci for support and the possibility to work on this interesting area. Additional financial support has been provided by the National Technology Agency of Finland (TEKES), which is gratefully acknowledged.

References

Aleksander, I., Dunmall, B. (2003). Axioms and Tests for the Presence of Minimal Consciousness in Agents. In O. Holland (Ed.), *Machine Consciousness* (pp. 7 - 18). UK: Imprint Academic.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, 22, 1999, 577-660.

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

Gregory, R. (2004). The blind leading the sighted. *Nature* Vol. 430, 19 August 2004 p. 836

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. (2005). 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.

Hesslow, G. (2001). *Medvetande som simulering av beteende och perception.* Retrieved Nov. 11, 2004, from http://www.mphy.lu.se/avd/nf/hesslow/philosophy/HW-medvet.htm

Hesslow, G. (2002). *Thinking as Simulation of Behaviour: an Associationist View of Cognitive Function.* Retrieved Nov. 11, 2004, from http://www.mphy.lu.se/avd/nf/hesslow/philosophy/ShortSimulation.htm

Hinton, G. E., McClelland, J. L., Rumelhart, D. E. (1990). Distributed Representations. In M. A. Boden (Ed), *The Philosophy of Artificial Intelligence*, 248 – 280. New York: Oxford University Press.

Holland, O., Goodman, R. (2003). Robots with Internal Models: A Route to Machine Consciousness? In O. Holland (Ed.), *Machine Consciousness* (pp. 77 - 109). UK: Imprint Academic.

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

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

Taylor, J. G. (1999). *The Race for Consciousness.* London, England: A Bradford Book, The MIT Press.

Thomas, N. J. T. (1999). Are Theories of Imagery Theories of Imagination? An Active Perception Approach to Conscious Mental Content. *Cognitive Science*, 23, 1999, 207-245

Consciousness, Emotion, and Imagination
A Brain-Inspired Architecture for Cognitive Robotics

Murray Shanahan
Dept. Electrical and Electronic Engineering
Imperial College London
Exhibition Road
London SW7 2BT
England
m.shanahan@imperial.ac.uk

Abstract

This paper proposes a brain-inspired cognitive architecture that incorporates approximations to the concepts of consciousness, emotion, and imagination. To emulate the empirically established cognitive efficacy of conscious as opposed to unconscious information processing in the mammalian brain, the architecture adopts a model of information flow from global workspace theory. Cognitive functions such as anticipation and planning are realised through internal simulation of interaction with the environment. Action selection, in both actual and internally simulated interaction with the environment, is mediated by affect. An implementation of the architecture is described which is based on weightless neurons and is used to control a simulated robot.

1 Introduction

From its inception to the present day, mainstream cognitive science has assumed language and reason to be the right conceptual foundations on which to build a scientific understanding of cognition. By contrast, the champions of biologically-inspired AI jetisoned these concepts in the 1990s. But at the same time they abandoned the very idea of cognition as a primary object of study. The present paper takes it for granted that understanding cognition will be central to achieving human-level artificial intelligence. However, the brain-inspired architecture described here, instead of manipulating declarative, language-like representations in the manner of classical AI, realises cognitive function through the animation of *analogical* (or *iconic*) representations whose structure is close to that of the sensory input of the robot whose actions they mediate (Sloman, 1971; Glasgow, *et al*., 1995).

Analogical representations are especially advantageous in the context of spatial cognition, which is a crucial capacity for any intelligent robot. While common sense inferences about shape and space are notoriously difficult with traditional logic-based approaches (Shanahan, 2004), in an analogical representation basic spatial properties such as distance, size, shape, and location are inherent in the medium itself and require negligible computation to extract. Furthermore, traditional language-like representations bear a subtle and contentious relationship to the world they are supposed to represent, and raise difficult questions about intentionality and symbol grounding (Harnad, 1990; Shanahan, 2005). With analogical representations, which closely resemble raw sensory input, this semantic gap is small and these questions are more easily answered.

In addition to these representational considerations, the design of the proposed architecture reflects the view, common among proponents of connectionism, that parallel computation should be embraced as a foundational concept rather than sidelined as a mere implementation issue. The present paper advocates a computational architecture based on the *global workspace* model of information flow, in which a serial procession of states emerges from the interaction of many separate, parallel processes (Baars, 1988; 2002). This *serial* procession of states, which includes the unfolding of conscious content in human working memory (Baars, & Franklin, 2003), facilitates anticipation and planning and enables a cognitively-enhanced form of action selection. Yet the robustness and flexibility of these cognitive functions depends on the behind-the-

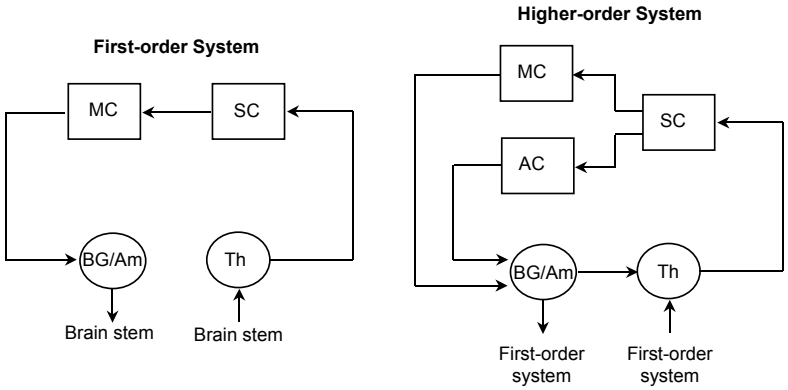


Fig. 1: A top-level schematic of the architecture. MC = motor cortex, SC = sensory cortex, AC = association cortex, BG = basal ganglia, Am = amygdala, Th = thalamus.

scenes performance of extremely large numbers of *parallel* computations, only the most relevant of which end up making a contribution to the ongoing serial thread (Shanahan & Baars, 2005).

The proposed architecture makes informal appeal to the concepts of consciousness, emotion, and imagination. Although only rough approximations to their humanly-applicable counterparts, the way these concepts are deployed here is inspired by their increasingly important role in the brain sciences (Damasio, 2000).

- **Consciousness** As already touched on, global workspace theory proposes a model of information flow in which conscious information processing is cognitively efficacious because it integrates the results of the brain's massively parallel computational resources (Baars, 1988; 2002). The theory has previously been used in the design of software agents (Franklin & Graesser, 1999), but is here applied to robotics for the first time.
- **Emotion** Based on clinical studies, Damasio (1995) argued persuasively that the human capacity for rational deliberation is dependent on an intact affective system, and many other cognitive scientists subscribe to the view that affect addresses the problems of decision making and action selection (Picard, 1997; Sloman, 2001). It permits a number of factors to be blended together and brought to bear on the problem of contention for resources (ie: muscles) by different brain processes. Neurologically plausible mechanisms of action selection compatible with this idea have already been demonstrated in a robotics setting (Prescott, *et al*; 1999; Cañamero, 2003).
- **Imagination** A number of neuroscientists have advanced the view that thought is internally simulated interaction with the environment or, to put it another way, the rehearsal of trajectories through sensorimotor space prior to their enactment (Cotterill, 1998; 2001; Hesslow, 2002). A small set of researchers have applied such ideas to robotics, including Chrisley (1990), Stein (1995), Holland (2003), and Hoffmann & Möller (2004).

The present architecture includes analogues of each of the following brain structures: the thalamus (for global broadcast), multiple motor-cortical populations (that compete for access to a global workspace), internal sensorimotor loops (capable of rehearsing trajectories through sensorimotor space), the basal ganglia (to carry out action selection), and the amygdala (to guide action selection through affect).

2 A Top-level Schematic

Fig. 1 shows a top-level schematic of the architecture. It can be thought of in terms of two interacting sub-systems. The first-order system is purely reactive, and determines an immediate motor response to the present situation without the intervention of cognition. But these unmediated motor responses are subject to a veto imposed by BG (the basal ganglia analogue). Through BG, which carries out salience-based action selection, the higher-order loop modulates the behaviour of the first-order system. It does this by adjusting the salience of currently executable actions. Sometimes this adjustment will have no effect. But sometimes it will result in a new action becoming the most salient. And sometimes it will boost an action's salience above the threshold required to release its veto, bringing about that action's execution.

The higher-order system computes these salience adjustments by carrying out off-line rehearsals of trajectories through (abstractions of) the robot's sensorimotor space. In this way – through the exer-

cise of its "imagination" – the robot is able to anticipate and plan for potential rewards and threats without exhibiting overt behaviour.

The first- and higher-order systems have the same basic components and structure. Both are sensorimotor loops. The key difference is that the first-order loop is closed through interaction with the world itself while the higher-order loop is closed internally. This internal closure is facilitated by AC, which simulates — or generates an abstraction of — the sensory stimulus expected to follow from a given motor output, and fulfils a similar role to a *forward model* in the work of various authors (Demiris & Hayes, 2002; Wolpert, *et al.*, 2003; Grush, 2004). The cortical components of the higher-order system (SC, AC, and MC) correspond neurologically to regions of association cortex, including the prefrontal cortex which is implicated in planning and working memory (Fuster, 1997).

2.1 Affect and Action Selection

Analogues of various sub-cortical and limbic structures appear in both the first- and higher-order systems, namely the basal ganglia, the amygdala, and the thalamus. In both systems, the basal ganglia are involved in action selection. Although, for ease of presentation, the schematic in Fig. 1 suggests that the final stage of motor output before the brain stem is the basal ganglia, the truth is more complicated in both the mammalian brain and the robot architecture it has inspired.

In the mammalian brain, the pertinent class of basal ganglia circuits originate in cortex, then traverse a number of nuclei of the basal ganglia, and finally pass through the thalamus on their way back to the cortical site from which they originated. The projections up to cortex are thought to effect action selection by suppressing all motor output except for that having the highest salience, which thereby makes it directly to the brain stem and causes muscular movement (Mink, 1996; Redgrave, *et al.*, 1999). The basolateral nuclei of the amygdala are believed to modulate the affect-based salience information used by the basal ganglia through the association of cortically mediated stimuli with threat or reward (Baxter & Murray, 2002; Cardinal, *et al.*, 2002).

The robot architecture includes analogues of the basal ganglia and amygdala that function in a similar way. These operate in both the first- and higher-order systems. In the first-order system, the amygdala analogue associates patterns of thalamocortical activation with either reward or punishment, and thereby modulates the salience attached to each currently executable action. The basal ganglia analogue adjudicates the competition between each executable action and, using a winner-takes-all strategy, selects the most salient for possible execution. While the salience of the selected action falls below a given threshold it is held on veto, but as soon as its salience exceeds that threshold it is executed.

The roles of the basal ganglia and amygdala analogues in the higher-order system are similar, but not identical, to their roles in the first-order system (Cotterill, 2001). These structures are again responsible for action selection. However, action selection in the higher-order system does not determine overt behaviour but rather selects one path through the robot's sensorimotor space for inner rehearsal in preference to all others. Moreover, as well as gating the output of motor association cortex (MC), the basal ganglia analogue must gate the output of sensory association cortex (AC) accordingly, and thus determine the next hypothetical sensory state to be processed by the higher-order loop.

This distinction between first-order and higher-order functions within the basal ganglia is reflected in the relevant neuroanatomy. Distinct parallel circuits operate at each level (Nolte, 2002, p. 271). In the first-order circuit, sensorimotor cortex projects to the putamen (a basal ganglia input nucleus), and then to the globus pallidus (a basal ganglia output nucleus), which projects to the ventral lateral and ventral anterior nuclei of the thalamus, which in turn project back to sensorimotor cortex. In the higher-order circuit, association cortex projects to the caudate nucleus (a basal ganglia input structure), and then to the substantia nigra (a basal ganglia output nucleus), which projects to the mediodorsal nucleus of the thalamus, which in turn projects back to association cortex.

2.2 Global Workspace Theory

Global workspace theory advances a model of information flow in which multiple, parallel, specialist processes compete and co-operate for access to a global workspace (Baars, 1988). Gaining access to the global workspace allows a winning coalition of processes to broadcast information back out to the entire set of specialists (Fig. 2). Although the global workspace exhibits a serial procession of broadcast states, each successive state itself is the integrated product of parallel processing.

According to global workspace theory, the mammalian brain instantiates this model of information flow, which permits a distinction to be drawn between conscious and unconscious information processing. Information that is broadcast via the global workspace is consciously processed while informa-

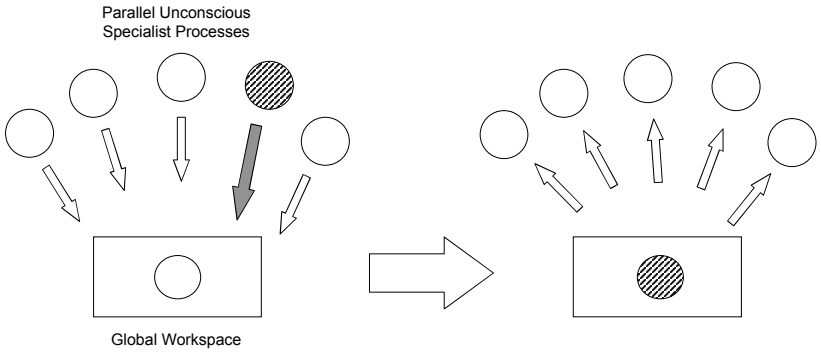


Fig. 2: The global workspace architecture.

tion processing that is confined to the specialists is unconscious. A considerable body of empirical evidence in favour of this distinction has accumulated in recent years (Baars, 2002).

The particular blend of serial and parallel computation favoured by global workspace theory suggests a way to address the frame problem – in the philosopher's sense of that term (Fodor, 2000) – which in turn suggests that conscious information processing may be cognitively efficacious in a way that unconscious information processing is not (Shanahan & Baars, 2005). In particular, in the context of so-called informationally unencapsulated cognitive processes, it allows relevant information to be sifted from the irrelevant without incurring an impossible computational burden. More generally, broadcast interleaved with selection facilitates the integration of the activities of large numbers of specialist processes working separately. So the global workspace model can be thought of as one way to manage the massively parallel computational resources that surely underpin human cognitive prowess.

The architecture of this paper conforms to the global workspace model of information flow by incorporating complementary mechanisms for the broadcast of information to multiple cortical areas and for selection between competing patterns of activation within those areas (Fig. 3). As Fig. 3 shows, the thalamus analogue is the locus of broadcast in the architecture. Information fans out from the thalamus to multiple cortical sites (within which it may be subject to further local distribution). Conversely, information funnels back into the thalamus, after competition within cortically localised regions, thanks to a process of selection between cortical sites realised by the basal ganglia.

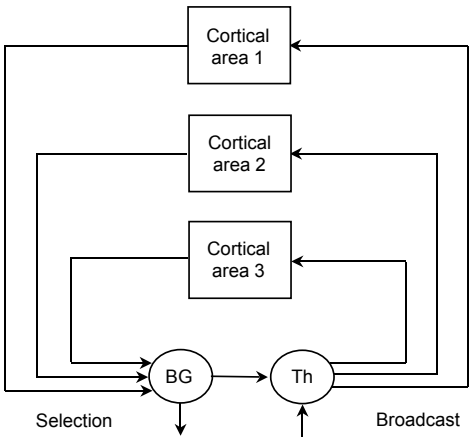


Fig 3: The fan-and-funnel model

This design reflects the fact that the first-order / higher-order distinction is preserved in the biological thalamus, which contains not only first-order relays that direct signals from the brain stem up to cortex (located, for example, in the lateral geniculate nucleus), but also higher-order relays that route cortical traffic back up to cortex (located, for example, in the pulvinar) (Sherman & Guillery, 2001; 2002). For this reason, and because of its favourable anatomical location and connectivity, the thalamus is a plausible candidate for a broadcast mechanism in the mammalian brain.

The fan-and-funnel model of broadcast / distribution and competition / selection can be straightforwardly combined with the top-level schematic of Fig. 1, as is apparent from the diagrams. Indeed, the role of the BG component of the higher-order loop introduced in Fig. 1 is precisely to effect a selection between the outputs of multiple competing cortical areas, as shown in Fig. 3.

3 An Implementation

The brain-inspired architecture of the previous section has been implemented using NRM (Dunmall, 2000), a tool for building large-scale neural network models using G-RAMs (generalising random access memories) (Figs. 4 and 5). These are weightless neurons employing single-shot training whose up-

date function can be rapidly computed (Aleksander, 1990).

The basic operation of a single G-RAM is illustrated in Fig. 4. The input vector is used to index a lookup table. In the example shown, the input vector of 1011 matches exactly with the fourth line of the table, which yields the output 6. When there is no exact match, the output is given by the line of the lookup table with the smallest Hamming distance from the input vector, so long as this exceeds a predefined threshold. In this example, if the input vector had been 1010, then none of the lines in the lookup table would yield an exact match. But the fourth line would again be the best match, with a Hamming distance of 1, so the output would again be 6. If no line of the lookup table yields a sufficiently close match to the input vector the neuron outputs 0, which represents quiescence.

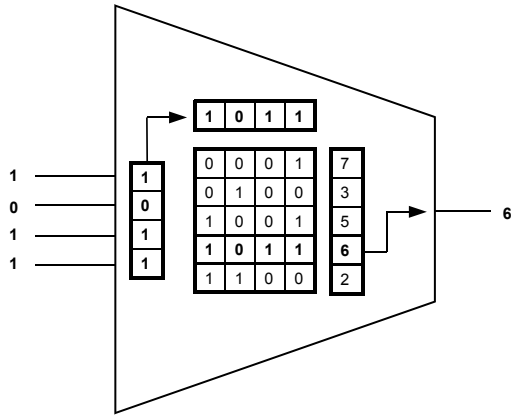


Fig 4: The G-RAM weightless neuron

The implemented system exploits the fact that G-RAMs can be easily organised into attractor networks with similar properties to Hopfield nets (Lockwood & Aleksander, 2003). The core of the implementation, which comprises almost 40,000 neurons and over 3,000,000 connections, is a set of cascaded attractor networks corresponding to each of the components identified in the architectural blueprint of the previous section.

The NRM model is interfaced to Webots, a commercial robot simulation environment (Michel, 2004). The simulated robot is a Khepera with a 64 × 64 pixel camera, and the simulated world contains cylindrical objects of various colours. The Khepera is programmed with a small suite of low-level actions including "rotate until an object is in the centre of the visual field" and "approach an object in the centre of the visual field". These two actions alone are sufficient to permit navigation in the robot's simple environment.

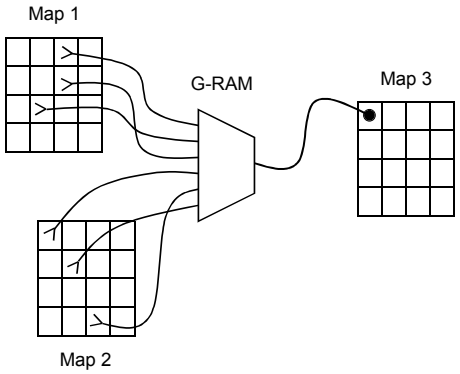


Fig 5: G-RAM maps and connections

The overall system can be divided into four separate modules – the visual system (Fig. 6), the affective system (Fig. 7), the action selection system (Fig. 8), and the broadcast / inner rehearsal system (Fig. 9). Each box in these figures denotes a layer of neurons and each path denotes a bundle of connections. If a path connects a layer A to an $n \times n$ layer B then it comprises n^2 separate pathways – one for each of the neurons in B – each of which itself consist of m input connections originating in a randomly assigned subset of the neurons in A (Fig. 5). For the majority of visual maps m is set to 32.

The two buffers in the visual system comprise 64 × 64 topographically organised neurons (Fig. 6). These are both attractor networks, a property indicated by the presence of a local feedback path. The transient buffer is activated by the presence of a new visual stimulus. The hallmark of a new stimulus is that it can jog the long-term visual buffer out of one attractor and into another. The higher-order thalamic relay of the inner rehearsal system is loaded from the transient visual buffer, whose contents rapidly fade allowing the dynamics of inner rehearsal to be temporarily dominated by intrinsic activity rather than sensory input.

The contents of the long-term visual buffer are fed to three competing motor-cortical areas, MC1 to MC3 (Fig. 8), each of which responds either with inactivity or with a recommended motor response to the current stimulus. Each recommended response has an associated salience (Fig. 7). This is used by the action selection system to determine the currently most salient action, which is loaded into the "selected action buffer" (Fig. 8). But the currently seleced action is subject to a veto. Only if its salience is sufficiently high does it get loaded into the "motor command" buffer, whose contents is forwarded to the robot's motor controllers for immediate execution.

So far the mechanism described is little different from a standard behaviour-based robot control ar-

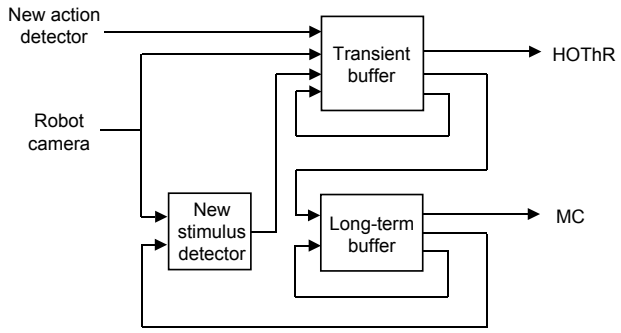


Fig. 6: Visual system circuitry (VC / IT). VC = visual cortex, IT = inferotemporal cortex.

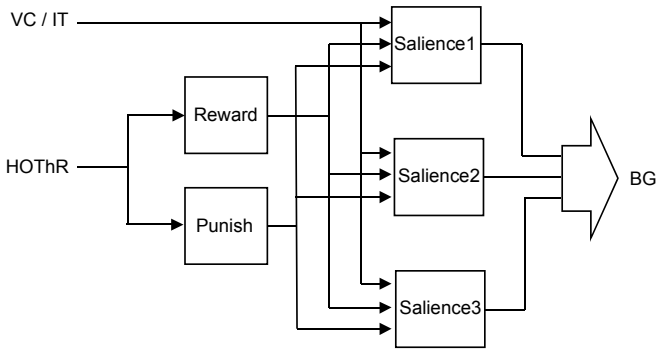


Fig. 7: Affect circuitry (Am)

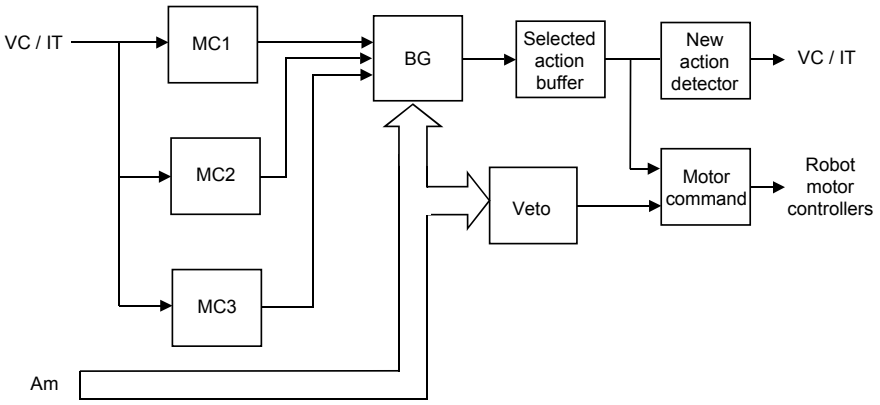


Fig. 8: Action selection circuitry (BG / MC)

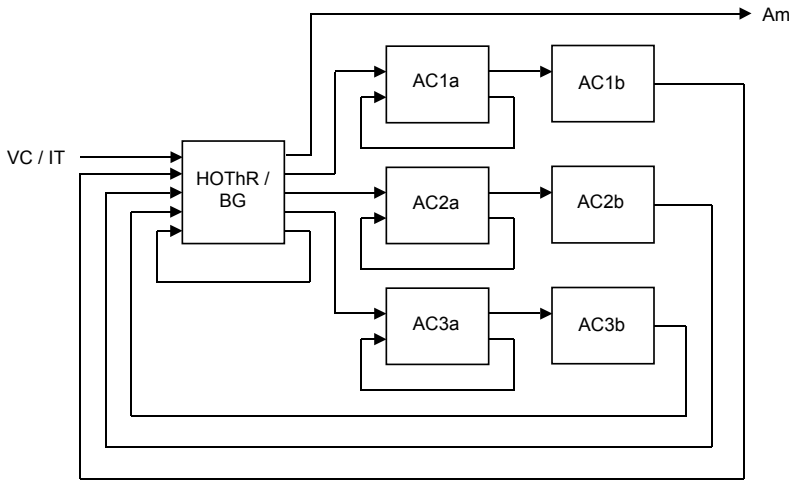


Fig. 9: Circuitry for broadcast and inner rehearsal (HOThR / BG / AC). HOThR = higher-order thalamic relay.

chitecture (Brooks, 1986). What sets it apart from a purely reactive system is its capacity for inner rehearsal. This is realised by the thalamocortical system depicted in Fig. 9. When a new visual stimulus arrives, it overwrites the present contents of HOThR, and is thereby broadcast to the three cortical association areas AC1a to AC3a. The contents of these areas stimulates the association areas AC1b to AC3b to take on patterns of activation corresponding to the expected outcomes of the actions recommended by their motor-cortical counterparts. These patterns are fed back to HOThR / BG, lead-

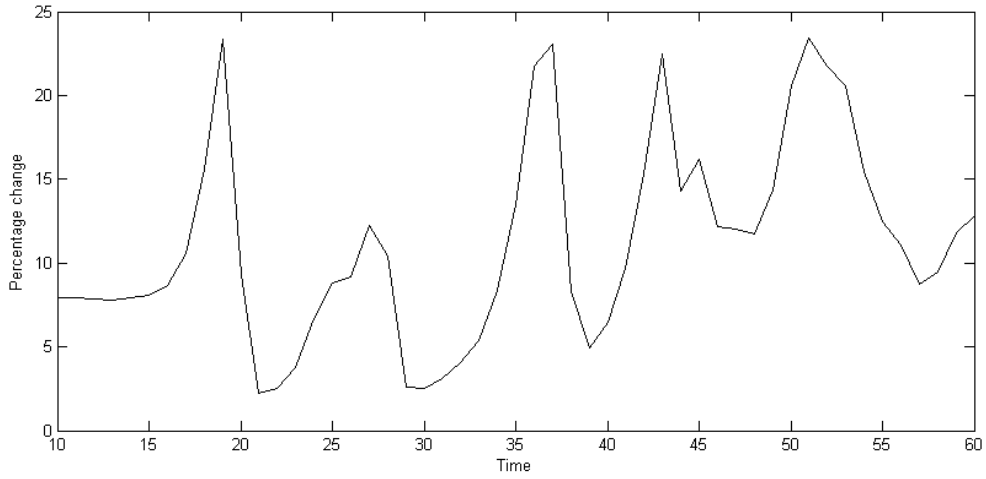


Fig. 10: Cycles of stability and instability in the thalamocortical system

ing to further associations corresponding to the outcomes of later hypothetical actions. By following chains of associations in this way, the system can explore the potential consequences of its actions prior to their performance, enabling it to anticipate and plan ahead.

But for this capacity to be useful, the system needs to be able to *evaluate* hypothetical futures as it discovers them. So as a result of inner rehearsal, the salience of the currently selected action becomes modulated according to the affective value of the situations to which it might lead (Fig. 7). If the currently selected action potentially leads to a desirable situation, a small population of "reward" neurons becomes active, causing an increase in the salience of that action. This in turn may be sufficient to trigger the release of its veto, bringing about its execution. Conversely, if the currently selected action potentially leads to an undesirable situation, a small population of "punish" neurons becomes active. The resulting decrease in salience of that action may cause a new action to become the most salient. In this case, the transient visual buffer is reloaded, its contents is passed on to HOThR, and the process of inner rehearsal is restarted.

4 Experimental Results

The implemented system currently runs on a 2.5 GHz Pentium 4 machine. Both Webots and NRM are run on the same machine, and the two systems communicate through an internal TCP socket. Under these somewhat unfavourable circumstances, each update cycle for the whole set of neurons takes approximately 750ms. A large proportion of this time is taken up by internal communication and graphics processing.

Fig. 10 illustrates an interesting property of the circuit of Fig. 9. The graph plots the percentage of neurons in the four maps HOThR and AC1a to AC3a that changed state from one time step to the next during a typical run in which no external sensory input was presented to the robot. (A similar pattern is typically produced soon after the initial presentation of an external stimulus.) The graph shows that the system of inner rehearsal exhibits a procession of stable states punctuated by episodes of instability, a pattern which is reminiscent of the phenomenon of aperiodic alternation between pancortical coherent and decoherent EEG activity reported by various authors (Rodriguez, *et al*., 1999; Freeman & Rogers, 2003).

The periods of stability depicted in the graph occur when the contents of HOThR is being successfully broadcast to the three cortical regions, while the spikes of instability indicate that HOThR is being nudged out of its previous attractor and is starting to fall into a new one. The new attractor will be the outcome of a competition between AC1b to AC3b. The resulting new contents of HOThR is then broadcast to AC1a to AC3a, causing new activation patterns to form in AC1b to AC3b, which in turn give rise to a renewed competition for access to HOThR. This tendency to chain a series of associations together is what gives the system its ability to look several actions ahead.

Table 1 presents an illustrative sequence of events that occurred in a typical run of the whole system in which this ability to look ahead is put to good use. The episode described starts with the initial presentation of a new stimulus to the robot's camera, and ends with the robot's first action. The time is given in perception-update-action cycles, so the overall time between stimulus and response is around 17 seconds. This suggests that real-time performance

would be attainable with current technology using a higher-end platform, assuming the Webots simulator is run on a different machine.

For the run presented, the robot's environment contained just three cylinders – one green, one red, and one blue. Area MC1 of the motor-cortical system was trained to recommend "rotate right" (RR) when presented with a green cylinder, while area MC2 was trained to recommend "rotate left" (RL). MC1's recommendation has the higher initial salience, and in a purely reactive system this action would be executed straight away. But thanks to the imposition of a veto, the inner rehearsal system gets a chance to anticipate the outcome of the recommended action. This turns out to be undesirable. So the system considers an alternative action, and this turns out to have a preferable expected outcome so is duly executed.

Table 1: An episode in a typical run

Time	Events
0	Green cylinder comes into view.
4	Green cylinder image in both visual buffers. MC1 recommends RR, MC2 recommends RL. RR has higher salience and is currently selected action. Veto is on.
7	Green cylinder image in HOThR and broadcast to AC1a to AC3a. AC1b has association with red cylinder, AC2b has association with blue cylinder.
8	Associated red cylinder image now in HOThR.
11	"Punish" neurons active, salience of RR going down.
13	Salience of RR very low. RL becomes currently selected action.
14	Transient visual buffer reloaded with green cylinder image.
16	Green cylinder image in HOThR and broadcast to AC1a to AC3a.
20	Associated blue cylinder image now in HOThR. "Reward" neurons active. Salience of RL going up.
22	Salience of RL very high. Veto released.
23	RL passed on to motor command area. Robot rotates left until blue cylinder in view.

5 Discussion

Although only a prototype, the implemented system has demonstrated the viability of the proposed architecture. As this episode illustrates, a system conforming to the architecture is capable of generating a cognitively enhanced motor response to an ongoing situation. The design methodology used is, of course, quite different to that currently favoured by researchers in mainstream cognitive robotics (Lespérance, *et al*., 1994), and is more closely allied to the research programme hinted at by Clark and Grush (1999). In place of viewpoint-free propositional representations, the present system employs viewer-centred analogical representations, and in place of symbolic reasoning it deploys a recurrent cascade of attractor networks. But compared with related products of the classical approach, the current implementation inherits certain several well-known disadvantages.

- While traditional propositional representations possess a compositional structure, and therefore comply with Fodor and Pylyshyn's *systematicity* constraint (Fodor & Pylyshyn, 1988), this is not true of the patterns of neuronal activity in the present system.
- Traditional propositional representations are adept at coping with *incomplete information* using disjunction and existential quantification. The present system can only deal with alternatives by using competitive parallelism and by exploring different threads of possibility at different times.
- Traditional planning systems are typically capable of effecting a *complete search* of the space of possible plans, while the presently implemented system of inner rehearsal ignores large tracts of search space and is only capable of a very crude form of backtracking.

Each of these issues is the subject of ongoing research. Brain-inspired cognitive architectures are relatively unexplored in artificial intelligence, and much work needs to be done before they can offer a viable alternative to the classical methodology in the domain of cognition.

But in addition to its potential engineering application, the architecture presented here can be construed as a concrete statement of a specific hypothesis about human brain function. In line with the methodological stance outlined in the paper's opening paragraph, this hypothesis ascribes the capacity for high-level cognition to the interplay of consciousness, emotion, and imagination. Building a computer model and using it to control a robot is one way to give a clear interpretation to these concepts and to make precise their hypothesised role in mediating behaviour.

To conclude, let's consider the extent to which these philosophically difficult concepts of consciousness, emotion, and imagination can legitimately be applied to artefacts that conform to the architectural blueprint of the present paper, such as

the implemented robot controller described in the previous section.

Let's begin with the concept of consciousness. The architecture respects all five of the "axioms of consciousness" proposed by Aleksander & Dunmall (2003). However, the present paper draws more heavily on the empirically grounded distinction between conscious and unconscious information processing hypothesised by global workspace theory (Baars, 1988; 2002). This carries over straightforwardly to the thalamocortical system of Fig. 9. The processing of activation patterns that appear in HOThR and are subsequently successfully broadcast to cortex can be considered "conscious", while all other information processing that goes on in the system is "unconscious". In accordance with global workspace theory, information that has been thus processed "consciously" integrates the contributions of many parallel processes, although the parallelism is very small-scale in the implemented robot controller described here.

Similar considerations apply to the concepts of emotion and imagination. The functional role of the affective and inner rehearsal systems in the present architecture is identical to that proposed for emotion and imagination by many authors for the human case (Damasio, 1995; 2000; Harris, 2000). The argument, in a nutshell, is that "human beings have evolved a planning system in which felt emotion plays a critical role. By imagining what we might do, we can trigger in an anticipatory fashion the emotions that we would feel were we to actually do it" (Harris, 2000, p. 88). In much the same vein, the higher-order loop of Fig. 9 "imagines" what the robot might do, and this triggers an "emotional" response in the affective system of Fig 7.

However, the liberal use of scare quotes in the above paragraphs remains appropriate. There are many possible objections to the literal application of concepts such as consciousness and emotion to a robot such as the one described here. Prominent among these is the sheer poverty of the robot's external environment, the consequent poverty of its control system's internal dynamics, and the limited range of behaviour it can exhibit as a result. But consider a future humanoid robot in an unconstrained natural environment, equipped with a control system conforming to the proposed architecture. Suppose the robot's broadcast / inner rehearsal system comprised not six cortical regions but 100,000. Perhaps it would be harder to rein in the use of these concepts in such a case. But for now this remains pure science fiction.

Acknowledgements

For help, discussion, and inspiration thanks to Igor Aleksander, Bernie Baars, Lola Cañamero, Ron Chrisley, Rodney Cotterill, Yiannis Demiris, Barry Dunmall, Ray Guillery, Gerry Hesslow, Owen Holland, Mercedes Lahnstein, Pete Redgrave, and S.Murray Sherman.

References

Aleksander, I. (1990). Neural Systems Engineering: Towards a Unified Design Discipline? *Computing and Control Engineering Journal 1*(6), 259–265.

Aleksander, I. & Dunmall, B. (2003). Axioms and Tests for the Presence of Minimal Consciousness in Agents. *Journal of Consciousness Studies 10* (4–5), 7–18.

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

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

Baars, B.J. & Franklin, S. (2003). How Conscious Experience and Working Memory Interact. *Trends in Cognitive Science 7* (4), 166–172.

Baxter, M.G. & Murray, E.A. The Amygdala and Reward. *Nature Reviews Neuroscience 3*, 563–573.

Brooks, R.A. (1986). A Robust Layered Control System for a Mobile Robot. *IEEE Journal of Robotics and Automation 2*, 14–23.

Cañamero, L.D. (2003). Designing Emotions for Activity Selection in Autonomous Agents. In R.Trapple, P.Petta & S.Payr (eds.), *Emotions in Humans and Artifacts*, MIT Press, pp. 115–148.

Cardinal, R.N., Parkinson, J.A., Hall, J. & Everitt, B.J. (2002). Emotion and Motivation: The Role of the Amygdala, Ventral Striatum, and Prefrontal Cortex. *Neuroscience and Biobehavioral Reviews 26*, 321–352.

Chrisley, R. (1990). Cognitive Map Construction and Use: A Parallel Distributed Processing Approach. In D.Touretzky, J.Elman, G.Hinton, and T.Sejnowski (eds.), *Connectionist Models: Proceedings of the 1990 Summer School*, Morgan Kaufman, pp. 287-302.

Clark, A. & Grush, R. (1999). Towards a Cognitive Robotics. *Adaptive Behavior 7* (1), 5–16.

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

Cotterill, R. (2001). Cooperation of the Basal Ganglia, Cerebellum, Sensory Cerebrum and Hippocampus: Possible Implications for Cognition, Consciousness, Intelligence and Creativity. *Progress in Neurobiology 64*, 1–33.

Damasio, A.R. (1995). *Descartes' Error: Emotion, Reason and the Human Brain*. Picador.

Damasio, A.R. (2000). *The Feeling of What Happens: Body, Emotion and the Making of Consciousness*. Vintage.

Demiris, Y. & Hayes, G. (2002). Imitation as a Dual-Route Process Featuring Predictive and Learning Components: a Biologically-Plausible Computational Model. In K.Dautenhahn & C.Nehaniv (eds.), *Imitation in Animals and Artifacts*, MIT Press, pp. 327–361.

Dunmall, B. (2000). *Representing the Sensed World in a Non-Biological Neural System*. Dept. Electrical & Electronic Engineering, Imperial College London.

Fodor, J.A. (2000). *The Mind Doesn't Work That Way*. MIT Press.

Fodor, J.A. & Pylyshyn, Z.W. (1988). Connectionism and Cognitive Architecture: A Critique. *Cognition 28*, 3–71.

Franklin, S. & Graesser, A. (1999). A Software Agent Model of Consciousness. *Consciousness and Cognition 8*, 285–301.

Freeman, W.J. & Rogers, L.J. (2003). A Neurobiological Theory of Meaning in Perception Part V: Multicortical Patterns of Phase Modulation in Gamma EEG. *International Journal of Bifurcation and Chaos* 13(10), 2867–2887.

Fuster, J.M. (1997). *The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe*. Lippincott-Raven.

Glasgow, J., Narayanan, N.H. & Chandrasekaran, B. (1995). *Diagrammatic Reasoning: Cognitive and Computational Perspectives*. MIT Press.

Grush, R. (2004). The Emulation Theory of Representation: Motor Control, Imagery, and Perception. *Behavioral and Brain Sciences*, in press.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D* 42: 335–346.

Harris, P.L. (2000). *The Work of the Imagination*. Blackwell.

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

Hoffmann, H. & Möller, R. (2004). Action Selection and Mental Transformation Based on a Chain of Forward Models. In *Proc. 8th International Conference on the Simulation of Behaviour (SAB 04)*, pp. 213–222.

Holland, O. (2003). Robots with Internal Models. *Journal of Consciousness Studies 10* (4–5), 77–109.

Lespérance, Y., Levesque, H.J., Lin, F., Marcu, D., Reiter, R. & Scherl, R.B. (1994). A logical approach to high-level robot programming: A progress report. In B.Kuipers (ed.), *Control of the Physical World by Intelligent Systems: Papers from the 1994 AAAI Fall Symposium*, pp. 79–85.

Lockwood, G.G. & Aleksander, I (2003). Predicting the Behaviour of G-RAM Networks. *Neural Networks 16*, 91–100.

Michel, O. (2004). Webots: Professional Mobile Robot Simulation. *International Journal of Advanced Robotics Systems 1* (1), 39–42.

Mink, J.W. (1996). The Basal Ganglia: Focused Selection and Inhibition of Competing Motor Programs. *Progress in Neurobiology 50*, 381–425.

Nolte, J. (2002). *The Human Brain: An Introduction to its Functional Anatomy*. Mosby.

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

Prescott, T.J., Redgrave, P. & Gurney, K. (1999). Layered Control Architectures in Robots and Vertebrates. *Adapative Behavior 7*, 99–127.

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

Rodriguez, E., George, N., Lachaux, J.-P., Martinerie, J., Renault, B. & Varela, F. (1999). Perception's Shadow: Long-Distance Synchronization of Human Brain Activity. *Nature* 397, 430–433.

Shanahan, M.P. (2004). An Attempt to Formalise a Non-Trivial Benchmark Problem in Common Sense Reasoning, *Artificial Intelligence 153*, 141–165.

Shanahan, M.P. (2005). Perception as Abduction: Turning Sensor Data into Meaningful Representation, *Cognitive Science 29*, 109–140.

Shanahan, M.P. & Baars, B. (2005). Applying Global Workspace Theory to the Frame Problem. *Cognition*, in press.

Sherman, S.M. & Guillery, R.W. (2001). *Exploring the Thalamus*. Academic Press.

Sherman, S.M. & Guillery, R.W. (2002). The Role of Thalamus in the Flow of Information to Cortex. *Philosophical Transactions of the Royal Society B* 357, 1695–1708.

Sloman, A. (1971). Interactions Between Philosophy and Artificial Intelligence: The Role of Intuition and Non-Logical Reasoning in Intelligence. *Artificial Intelligence 2*, 209–225.

Sloman, A. (2001). Beyond Shallow Models of Emotion. *Cognitive Processing 2* (1), 177–198.

Stein, L.A. (1995). Imagination and Situated Cognition. In K.M.Ford, C.Glymour & P.J.Hayes (eds.), *Android Epistemology*, MIT Press, pp. 167–182.

Wolpert, D.M., Doya, K. & Kawato, M. (2003). A Unifying Computational Framework for Motor Control and Social Interaction. *Philosophical Transactions of the Royal Society B* (358), 593–602.

Chaotic Itinerancy, Active Perception and Mental Imagery

Takashi Ikegami
*Department of General Systems Sciences,
The Graduate School of Arts and Sciences, University of Tokyo
3-8-1 Komaba, Tokyo 153-8902, Japan
ikeg@sacral.c.u-tokyo.ac.jp

Abstract

A new theory of perception in terms of scale and time ordering is proposed. A simulated agent is convoluted with an interconnected FitzHugh-Nagumo neuron networks via pulse signaling. An agent senses the world through the input neurons and computes the motor outputs by the internal network. So-called chaotic itinerant behavior of an agent demonstrates that perception is associated with self-navigated active motions.

1 introduction

Consciousness concerns self-organized temporal structures in one's brain. The approach of traditional dynamical systems has been studying these temporal structures in terms of flows in certain state spaces. For example, chaotic itinerancy (CI) was found to be a flow pattern unique to many high-dimensional dynamical systems (Kaneko and I., 2003). CI is a spontaneous itinerating among local, typically chaotic, attractors. Such system demonstrates chaotic dynamics but eventually transits to chaotic dynamics of different kinds.

However, CI is ubiquitous in the high dimensional systems, irrespective of living or non-living states. Therefore, CI isn't sufficient for characterizing the living state. In order to discuss the internal process of the living systems, such as conscious states, we must invent a new logic/dynamics additional to the idea of CI (Tsuda, 2001; Tani, 1998).

In this paper, the investigation in terms of CI will be made on the dynamic nature of perception states associated with self-motion behavior, i.e. active perception (Gibson, 1962). Active perception (AP) isn't a simple interplay between perception and self-motion, which involves more complex processing such as exploration and bundling of action patterns. Specifically, a single action involves unseen action patterns, which results in some explorative behaviors. Active perception insists that exploration process organizes our sensory experience, stressing that the process of exploration itself is more important than its attainment.

2 Modeling with FHN network

Combining the ideas of CI and AP, I present a new perception theory and its realization as a computational model. This model involves a mobile agent that is endowed a network of the FitzHugh-Nagumo neurons (Fitzhugh, 1961; Nagumo and Yoshizawa, 1962). The FitzHugh-Nagumo(FHN) model is a simplification of the Hodgkin-Huxley (1952) model. Each FHN neuron consists of two variables, a 'fast' variable (u) corresponding to the membrane potential and the 'slow' variable (v).

$$\frac{du}{dt} = c(u - \frac{u^3}{3} - v + I(t)) \tag{1}$$

$$\frac{dv}{dt} = a + u - bv \tag{2}$$

Here the input signal $I(t)$ is posed on the fast variable u. The system has been studied intensively its bifurcation structure (see, e.g. the review by (Kostova and Schonbek, 2004)). Recently it has shown that a spiking behavior of periodic and chaotic inter-spike intervals are parameterized by the periodicity of the pulse trains of $I(t)$. Chaotic behavior is caused by the sub-threshold dynamics, in particular(et al., 1996).

In the present work, we study a network of the FHN neurons with the time delayed connections. For example, when a fast variable of neuron is activated ($u > 0$), a pulse signal is transmitted to the connected neurons with a time delay. The connected neuron k will receive a pulse signal of the width w and the height I after the time-lag of $\tau(n)$. This quantity $\tau(n)$

is a characteristic of a neuron n that emits the signal. Here simply, each neuron can either take τ_1 or τ_2.

An agent is assumed to have a circular body of a radius $R = 10$. The neurons are grouped into 3 categories; input, internal and output neurons. Those neurons are randomly and sparsely (e.g. the connection probability is set at 20%) connected with each other. In the present example, we use 10 input neurons, 16 internal neurons and 4 output neurons. The input neurons are arranged along the circumference at an equal distance. An environment is a two -dimensional pixel array with a graded amplitude. They receive signals from the pixel bit on its foot. If an agent stays in the same position, the input neuron receives a constant input from the pixel.

The output neurons cooperatively constitute the motor actions. When a output neuron m is activated, it produces a pulse output $I_m(t)$ of the width w_m and the height I_m. Motion of the agent is controlled by the two forward forces, $F_L(t)$ and $F_R(t)$. Here, each force is computed from two output neurons, respectively. We compute the forces at time t from the output neural states as follows;

$$F_{L/R}(t) = g \tanh(\sum_{n=L/R} I_n(t)). \qquad (3)$$

Using the forces, we also compute the rotating motion. Hence the agent's behavior is the combination of the forwarding and rotating motion. The internal neurons mediate the input and output neurons by mutually sending pulse trains.

In order to focus on the issue of time structure, we discarded some important aspects of generic neural networks. First, no excitatory or inhibitory signals are assumed. Due to the basic characteristics of the FHN neuron, pulse train inputs can activate a recipient neuron but also suppress it, depending on the width, height and periodicity of the input. Second, we don't integrate the signal amplitudes at the recipient neuron. Therefore, the exact coincident signals are equivalent to a single pulse. However, a slight difference in signaling timing will cause the irregular pulse train. For example, almost coincident signals are concatenated to produce a large pulse width. This happens to activate the neuron state.

3 Model behavior

I present an example of the agent's exploration pattern in a two-dimensional plane. In general, an agent is sensitive to the spatial figure. The example has a checkerboard pattern with two different strengths (I = 0.21 and 0.28) and some noisy scatters of strength (I = 0.07). In this parameter range, input neurons can't activate spontaneously. It only activates against the certain types of temporal input sequences.

An example of CI is depicted in Fig.1 and Fig.2. I take a sum of the internal neural states (INS) as an index for distinguishing different local attractors.

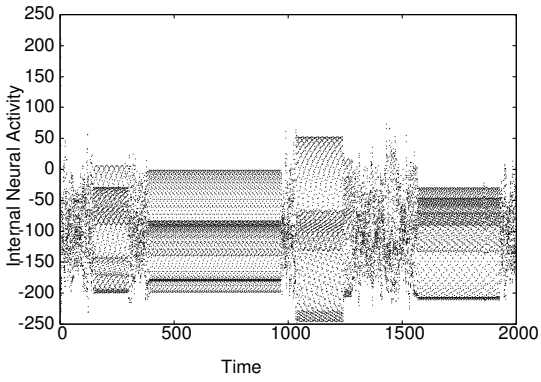


Figure 1: Itinerant behavior of the integrated internal neural activities. Different temporal pattern corresponds to the different local attractor, which is associated with the spatial patterns in Fig.2. The set of parameters are, $width = 0.1, I = 0.7, a = 0.7, b = 0.8$ and $C = 10$.

The present experiments suggest the following:

i) A variety of exploration dynamics is evident. Even for a fixed figure on the plane, an agent switches from one navigation style to another by smoothly touching the figure. The switching behavior from onr navigation style to the others is well correlated with the CI of the internal neural activity.

ii) The attractors that are triggered by a figure on the plane respond to the scanning speed of the sensors and the spatial scale of the figure. Namely, spatial scales are translated into time scales of the internal stimuli. As the result, the internal network tunes its navigation style on the plane, which in turn determines the succeeding input patterns.

iii) Regularity of the figure on the plane causes either totally chaotic or halting dynamics, except for some moderate spatial scales.

The above observations suggest that an agent doesn't simply respond to the physical stimuli. But it is responding to the *time structure of the input stimuli.* Namely, an agent perceives the world not based on the snapshot of "sense-data" but on the temporal flow of the "sense-data". Exploratory behavior with a certain style of motion produces the effective pulse trains, which cause the input neurons to fire and al-

ternate the successive exploratory behavior. On the other hand, the internal neurons are mutually sending pulse trains to each other. The above observations show that such internal state has a variety of local and global attractors. As a result of interference between the internal and input neurons, an agent autonomously selects the style of exploratory behavior.

4 Discussions

Henri Begson's notion of mental imagery (Bergson, 1911) can be seen in the interplay between motion and internal chaotic dynamics. His idea was to consider perception as a process based on the motion structure. Gibson implicitly inherited the idea and proposed a theory of active perception. Active perception considers that perception isn't caused by the momentary data of sense but by the successive stimuli of sensation. Exploration isn't executed inside of a memory space but in the real space. It is automatically guided by the consecutive sensory flow from the outside (Gibson, 1966).

Muenzinger and Tolman's vicarious try-and-error (Tolman, 1948; Muenzinger, 1938) or the more recent micro choices experiments of Brown (1992) show that normal rats demonstrate 'private' simulation or pantomime-like behaviors in some maze finding tasks. Oppose to the recent discussion of 'cognitive map' (memory-based) navigation, I insist that these results are empirical evidence of environment guided navigation. Exploration and selection of action is automatically guided by the rats' acquired disposition. In other words, the local layout of the maze and the way a landscape changes when a rat moves around provide a set of parameters that determine the disposition.

A dynamics that underlies the active perception is what I here clarified in terms of chaotic itinerancy with embodiment. As we saw in this paper, the time structure of the input stimuli, not the input itself, organizes the perceptual input, which coordinates the internal neural dynamics. And the time structure itself is generated by the self-motion pattern. We think that the conscious state as distinct from the material self can be studied explicitly by considering the hierarchy of time structure generated by an agent's self-motion.

Acknowledgments

This work is partially supported by Grant-in aid (No. 15300086) from the Ministry of Education, Science, Sports and Culture, The 21st Century COE (Center of Excellence) program(Research Center for Integrated Science) of the Ministry of Education, Culture, Sports, Science, and Technology, Japan, and the ECAgent project, sponsored by the Future and Emerging Technologies program of the European Community (IST-1940).

References

H. Bergson. *Matter and Memory*. London: George Allen and Unwin, 1911.

M.F. Brown. Does a cognitive map guide choices in a radial-arm maze? *J. Exp. Psychology*, 18(1):56–66, 1992.

Kaplan et al. Suthreshold dynamics in periodically stimulated squid giant axons. *Physical Review Letters*, 76(21):4074–4077, 1996.

R. Fitzhugh. Impulses and psychologial states in theoretical models of nerve membrane. *BioPhys.Journal*, 1(1):445–466, 1961.

J.J. Gibson. Observations on active touch. *Psychological Review*, 69:477–491, 1962.

J.J. Gibson. The problem of temporal order in stimulation and perception. *Journal of Psychology*, 62: 141–149, 1966.

K. Kaneko and Tsuda. I. Chaotic itinerancy. *CHAOS*, 13(3):926–936, 2003.

R. Kostova, T. Ravindran and M. Schonbek. Fitzhugh-nagumo revisited: Types of bifurcations, periodic forcing and stability regions by a lyapunov functional. *Int'l J. Bifurcation and Chaos*, 14(3):913–926, 2004.

K.F. Muenzinger. Vicarious trial and error at a point of choice, i: A general survey of its realization to learning efficiency. *J. Genet. Psychology*, 53:75–86, 1938.

S. Nagumo, J. Arimoto and S. Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, pages 2061–2070, 1962.

J. Tani. An interpretation of the 'self' from the dynamical systems perspective: A constructivist approach. *Journal of Consciousness Studies*, 5(56): 516–542, 1998.

E.E. Tolman. Cognitive maps in rats and men. *Psychol. Rev*, 55:189–208, 1948.

I. Tsuda. Toward an interpretation of dynamic neural activity in terms of chaotic dynamical systems. *Behav. Brain Sci.*, 24:575–628, 2001.

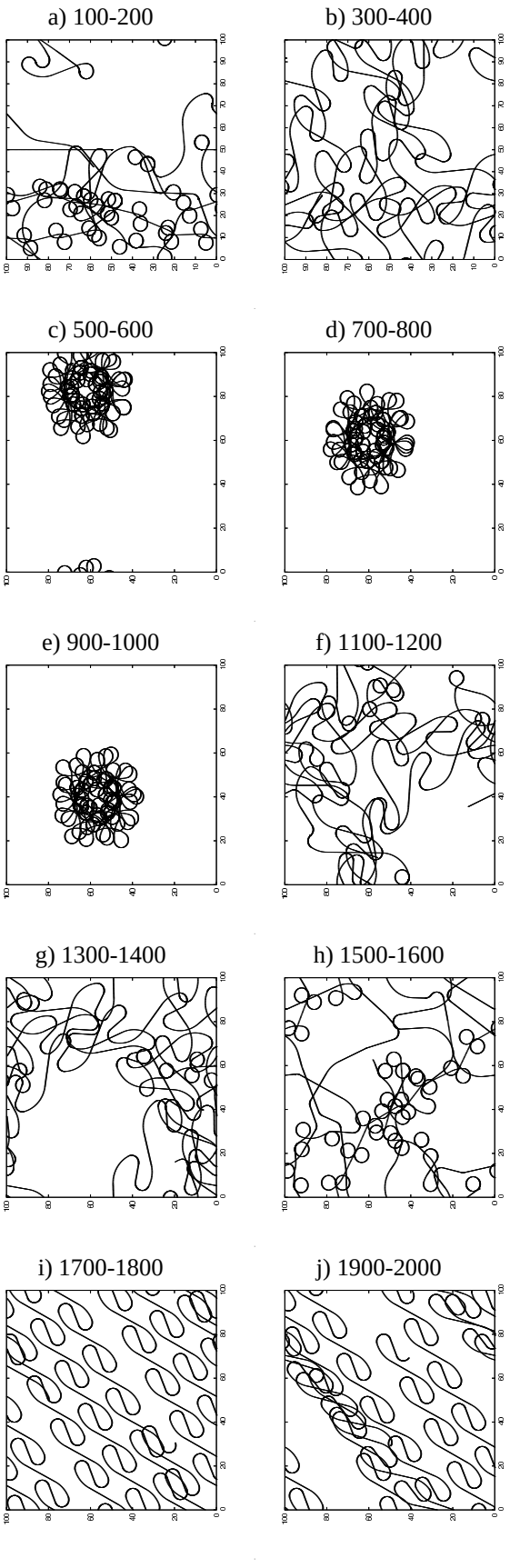


Figure 2: Time Evolution of Spatial Trails. A structure of motion pattern switches from one to the other. From the left upper corner (a) to the right bottom corner (j), spatial trails are overlaid for each 100 time steps for every other 100 time step.

Planning by imagination in Cicerobot, a robot for museum tours

Antonio Chella
University of Palermo
Viale delle Scienze, Palermo, Italy
chella@unipa.it

Marcello Frixione
University of Salerno
Via P. D. Melillo, Fisciano (SA), Italy
frix@dist.unige.it

Salvatore Gaglio
University of Palermo
Viale delle Scienze, Palermo, Italy
gaglio@unipa.it

Abstract

In the course of the years, authors developed a cognitive architecture for robot vision and action. One of the main characteristics of the architecture is the principled integration of perception and of symbolic knowledge by means of the introduction of an intermediate representation based on conceptual spaces. The proposed architecture may support artificial consciousness in the sense of Aleksander (1996). The paper describes in details how the proposed framework allows the robot to deliberate its own sequences of actions by means of planning and imagination cycle. In this perspective, planning is performed by taking advantage from the representations in conceptual space.

1 Introduction

The current generation of autonomous robots has showed impressive performances in mechanics and control of movements, see for instance the ASIMO robot by Honda or the QRIO by Sony. However, these state-of the-art robots present only limited capabilities to perceive, reason and act in a new and unstructured environment.

We claim that a new generation of autonomous robots, effectively able to perceive and act in unstructured environments and to interact with people, should be aware of their external and inner perceptions, should be able to pay attention to the relevant entities in their environment, to image, predict and to effectively plan their actions. In a word, they should include some form of artificial consciousness, in the sense of (Aleksander 1996).

In recent years, there has been an increasing interest towards computational models of "machine consciousness", see (Holland 2003) for a review. In the course of the years, we developed a cognitive architecture for robot vision (Chella et al. 1997, 2000). The architecture is experimented on an autonomous robot platform based on a RWI B21 robot equipped with a pan-tilt stereo head, laser rangefinder and sonars (Fig. 1). The aim of the architecture is to integrate visual perception with knowledge representation to generate conscious behaviour in a robot. One of the main characteristics is the principled integration of perception and of symbolic knowledge by means of the introduction of an intermediate representation based on conceptual spaces (Gärdenfors 2000).

We claim that the proposed architecture supports robot perception, attention, imagination, planning, emotions and sense of self; in other words, as explained in the rest of the paper, the architecture has all the means for artificial consciousness (Aleksander and Dunmall, 2003).

Figure 1: The RWI B21 robot.

In order to test the system in non trivial tasks, we employed our architecture in the Cicerobot project. The aim of the project is an autonomous robot able

to offer guided tours in our Department of Computer Engineering of the University of Palermo and at the Archaelogical Museum of Agrigento. The task is a significant case study for machine consciousness because it concerns perception, self perception, planning and human-robot interaction.

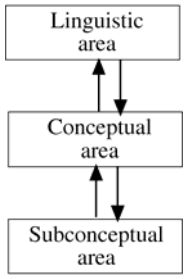


Figure 2: The three computational areas

2 The architecture of the robot

The proposed architecture (Fig. 2) is organized in three computational "areas", a term which is reminiscent of the cortical areas in the brain. The subconceptual area is mainly concerned with the processing of data coming from the sensors. Here information is not yet organized in terms of conscious structures and categories. From the point of view of the artificial vision, this area includes the processes that extract the 3D model of the perceived scene.

In the linguistic area, representation and processing are based on a logic-oriented formalism. The conceptual area is intermediate between the subconceptual and the linguistic areas. Here, data is organized in conceptual structures independent of any linguistic description. According to the "intermediate level theory of consciousness" proposed by Jackendoff (1987), this is the area where robot consciousness arises, because in this area the results of the "unconscious" low-level and thought procedures are represented.

In our model, the three areas are concurrent computational components working together on different commitments. There is no privileged direction in the flow of information among them: some computations are strictly bottom-up, with data flowing from the subconceptual up to the linguistic through the conceptual area; other computations combine top-down with bottom-up processing.

2.1 Conceptual Spaces

The conceptual area, as previously stated, is the intermediate area between the subconceptual and the linguistic area and it is the area where visual awareness arises, because, as previously stated, in this area the results of the low-level and unconscious thought procedures are stored and represented. This area is based on the theory of conceptual spaces (Gärdenfors 2000).

Conceptual spaces provide a principled way for relating high level, linguistic formalisms with low level, unstructured representation of data. A conceptual space CS is a metric space whose dimensions are related to the quantities processed in the subconceptual area. Different cognitive tasks can presuppose different conceptual spaces, and different conceptual spaces can be characterised by different dimensions.

Examples of possible dimensions are colour, pitch, mass, spatial coordinates, and so on. In general, dimensions are strictly related to the results of measurements obtained by sensors. In any case, dimensions do not depend on any specific linguistic description. In this sense, conceptual spaces come before any symbolic of propositional characterisation of cognitive phenomena.

We use the term c-knoxel to denote a point in a conceptual space. The term c-knoxel (in analogy with the term pixel) stresses the fact that a point in CS is the primitive element of robot artificial consciousness and, at the same time, the knowledge primitive element at the considered level of analysis.

The conceptual space CS acts as a workspace in which low-level and high-level processes access and exchange information respectively from bottom to top and from top to bottom, in agreement with the Global Workspace Theory (Baars 1988), (Dehaene and Naccache 2001). However, our conceptual space is a workspace with a precise geometric structure of metric space and also the operations in CS are geometrics: this structure allow us to describe the functionalities of the robot awareness in terms of the language of geometry.

It has been questioned if visual awareness is based on a 3D representation, as presupposed by Marr (Marr 1982) or a 2 ½ D representation as proposed by Jackendoff, or a combination of 2D views as proposed by (Ullman 1996). In the present architecture, we maintain the Marrian approach, according to which our c-knoxel corresponds to a moving 3D shape.

According with the hypothesis of asynchrony of consciousness (Zeki and Bartels 1998), the c-knoxel parameters may be considered the outcomes of asynchronous "microconsciousness" sites at the subsymbolic level. The role of the conceptual space is therefore to merge these outcomes in an unified entity, i.e., the c-knoxel. Aleksander and Dunmall

(2000) postulate that the correct relationships of the outcomes from the microconsciousness sites is held by j-referents, i.e., by the ego referents of the robot. In our architecture, the j-referents are implicit in the structure of the c-knoxel, because its parameters hold both information about object shape and about object position.

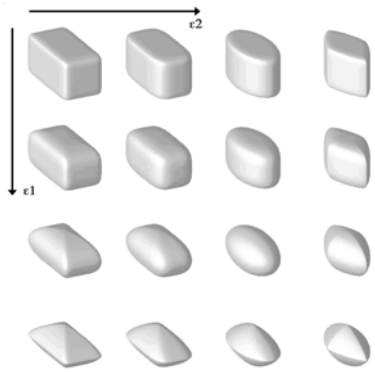


Figure 3: Superquadric shapes obtained by changing the form factors.

2.2 Object and Scene Representation

In (Chella et al. 1997) we assumed that, in the case of static scenes, a c-knoxel **k** coincides with a 3D primitive shape, characterised according to some constructive solid geometry (CSG) schema. In particular, we adopted superquadrics (Jaklič et at. 2000) as the primitive of CSG. Superquadrics allow us to deal with a compact description of the objects in the perceived scene. This approach is an acceptable compromise between the compression of information in the scene and the necessary computational costs Moreover, superquadrics provide good expressive power and representational adequacy.

Superquadrics are geometric shapes derived from the quadric parametric equation with the trigonometric functions raised to two real exponents. Fig. 3 shows the shape of a superquadric obtained by changing its form factors.

In the current implementation, apart from the parameters related with the shape and the position in space of the perceived superquadric, other c-knoxel parameters describe the "valuations" associated with the c-knoxel itself. Valuations are related with the feelings associated with c-knoxels, in the sense of (Jackendoff 1996), i.e., if the perceived entity is external or imagined, if it has an affective content, and so on. Some examples of valuations will be discussed in the rest of the paper.

In order to represent composite objects that cannot be reduced to single c-knoxels, we assume that they correspond to groups of c-knoxels in CS. For example, a chair can be naturally described as the set of its constituents, i.e., its legs, its seat and so on. Fig. 4 (left) shows a hammer composed by two superquadrics, corresponding to its handle and to its head.

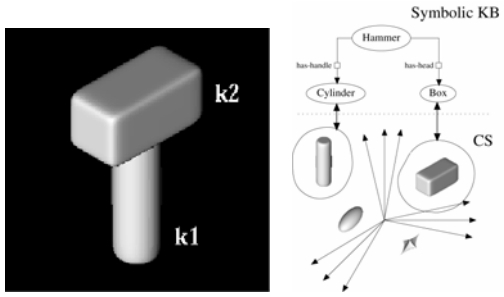


Figure 4: A hammer made up by two superquadrics and its representation in the conceptual space.

Fig. 4 (right) shows a picture of how hammers are represented in CS. The concept hammer consists of a set of pairs, each of them is made up of the two components of a specific hammer, i.e., its handle and its head.

2.3 Dynamic scenes

In order to account for the perception of dynamic scenes, we choose to adopt an intrinsically dynamic conceptual space. It has been hypothesised that simple motions are categorised in their wholeness, and not as sequences of static frames. In other words, we assume that simple motions of geometrically primitive shapes are our perceptual primitives for motion perception.

In our dynamic conceptual space, a c-knoxel now corresponds to a generalised simple motion of a superquadric. By generalised we mean that the motion can be decomposed in a set of components each of them associated with a degree of freedom of the moving superquadric.

A way of doing this, is suggested by the well known Discrete Fourier Transform (DFT), (see, e.g., Oppenheim and Shafer 1989). Given a parameter of the superquadric, e.g., a_x, consider the function of time $a_x(t)$; this function can be seen as the superimposition of a discrete number of trigonometric functions. This allows the representation of $a_x(t)$ in a discrete functional space, whose basis functions are trigonometric functions.

By a suitable composition of the time functions of all superquadric parameters, the overall function of time describing superquadrics parameters may be represented in its turn in a discrete functional space.

We adopt the resulting functional space as our dynamic conceptual space. This new CS can be taught as an "explosion" of the space in which each main axis is split in a number of new axes, each one corresponding to a harmonic component. In this way, a point **k** in the CS now represents a superquadric along with its own simple motion. This new CS is also consistent with the static space: a quiet superquadric will have its harmonic components equal to zero.

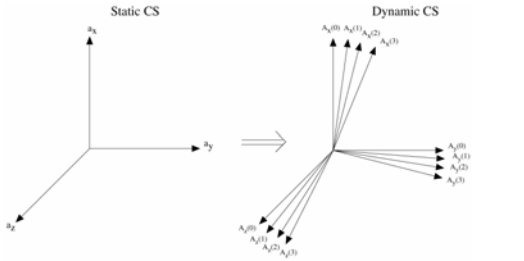


Figure 5: An evocative, pictorial representation of the static and dynamic conceptual spaces.

In Fig. 5 (left) a static CS is schematically depicted; Fig. 5 (right) shows the dynamic CS obtained from it. In the CS on the left, axes represent superquadric parameters; in the rightmost figure each of them is split in the group of axes, that represent the harmonics of the corresponding superquadric parameter.

2.4 Situations and Actions

Let us consider a scene made up by the robot itself along with other entities, like objects and persons. Entities may be approximated by one or more superquadrics. Consider the robot moving near an object. We call Situation this kind of scene. It may be represented in CS by the set of the c-knoxels corresponding to the simple motions of its components, as in Fig. 6 (left) where $\mathbf{k_a}$ corresponds to an obstace object, and $\mathbf{k_b}$ corresponds to the moving robot.

A Situation is therefore a configuration of c-knoxels that describe a state of affairs perceived by the robot. We can also generalize this concept, by considering that a configuration in CS may also correspond to a scene imagined or remembered by the robot. For example, a suitable imagined Situation may correspond to a goal, or to some dangerous state of affairs, that the robot must figure out in order to avoid it. Following the suggestion of (Jackendoff 1987), we added a new binary valuation that distinguish if the c-knoxel is effectively perceived, or it is imagined by the robot. In this way, the robot represents both its perceptions and its imaginations in conceptual space.

In a perceived or imagined Situation, the motions in the scene occur simultaneously, i.e., they correspond to a single configuration of c-knoxels in the conceptual space. To consider a composition of several motions arranged according to a temporal sequence, we introduce the notion of Action: an Action corresponds to a "scattering" from one Situation to another Situation in the conceptual space, as in Fig. 6 (right).

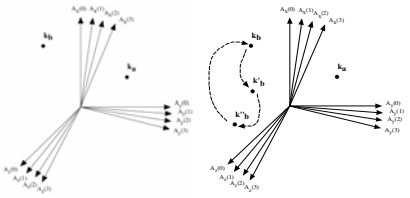


Figure 6: An example of Situation and Action in CS.

We assume that the situations within an action are separated by instantaneous events. In the transition between two subsequent configurations, a "scattering" of some c-knoxels occur. This corresponds to a discontinuity in time that is associated to an instantaneous event.

The robot may perceive an Action passively when it sees some changes in the scene, e.g., a person in the robot environment changing his/her position. More important, the robot may be the actor of the action itself, when it moves or when it interacts with the environment, e.g., when it pushes an object. In both cases, an Action corresponds to a transition from a Situation to another.

2.5 Linguistic Area

The representation of situations and actions in the linguistic area is based on a high level, logic oriented formalism. The linguistic area acts as a sort of "long term memory", in the sense that it is a semantic network of symbols and their relationships related with the robot perceptions and actions. The linguistic area also performs inferences of symbolic nature.

In the current implementation, the linguistic area is based on OpenCyc , an open source version of the knowledge base CYC (Lenat 1995), a hybrid KB in the KL-ONE tradition. A hybrid formalism in this sense is constituted by two different components: a terminological component for the description of concepts, and an assertional component, that stores information concerning a specific context.

In the domain of robot actions, the terminological component contains the description of relevant concepts such as `Situation`, `Action`, `Time_instant`, and so on.

In general, we assume that the description of the concepts in the symbolic KB is not completely exhaustive. We symbolically represent only that information that is necessary for inferences.

The assertional component contains facts expressed as assertions in a predicative language, in which the concepts of the terminological components correspond to one argument predicates, and the roles (e.g. `precond`, `part_of`) correspond to two argument relations.

The user may performs queries by using the symbolic language in order to orient the actions, for example, the user may ask the robot to search for an object.

Moreover, the system may generate assertions describing the robot current state, its perceptions, its planned actions, and so on. In this sense, the operations in the linguistic area acts as a generator of the "flow of consciousness" described by (Dennett 1993).

However, differently from Dennett, the terms in our linguistic area are strictly "anchored" to c-knoxels in the conceptual area, in the sense that the meaning of the terms in the linguistic area is represented by means of the corresponding c-knoxels in the conceptual area. Therefore, in our architecture, symbolic terms are strictly related with the robot visual awareness. Consciousness is not generated by the robot language, but, instead, the role of language is to "summarize" the dynamics of the c-knoxels at the conscious conceptual area.

Another role of the linguistic area is "help the robot to think" in Jackendoff terms, in the sense that the linguistic area merges the perceptual information coming from the CS with the facts stored in the KB in order to orient the computational resources of the robot.

3 Planning by imagination

The proposed framework for the interpretation of perceived robot situations and actions may be extended to allow the robot to deliberate its own sequences of actions. In this perspective, planning may be performed by taking advantage from the representations in CS. Note that we are not claiming that all kinds of planning must be performed within CS, but the forms of planning that are more directly related to perceptual information can take great advantage from visual awareness in the conceptual area.

In facts, the preconditions of an action can be simply verified by geometric inspections in the CS, while in the STRIPS planner (Fikes and Nilsson 1971) the preconditions are verified by means of logical inferences on symbolic assertions. Also the effects of an action are not described by adding or deleting symbolic assertions, as in STRIPS, but they can be easily described by the Situation resulting from the expectations of the execution of the action itself in CS.

In the proposed architecture, the recognition in a scene of a certain component of a Situation (a c-knoxel in CS) elicits the expectation of the other components of the Situation itself. The recognition of a certain Situation could also elicit the expectation of a scattering in the arrangement of the c-knoxels in the scene: the expectation mechanism generates the expectations for a different Situation in a subsequent CS configuration.

We take into account two main sources of expectations. On the one side, expectations are generated on the basis of the structural information stored in the symbolic knowledge base described in the previous Sect. We call linguistic such expectations. As soon as a Situation is recognized, which is the precondition of a certain Action, then the symbolic description elicit the expectation of the effect Situation.

On the other side, expectations could also be generated by a purely Hebbian association between situations. Suppose that the robot has learnt that when it sees somebody pointing on the right, it must turn in that direction. The system learns to associate these situations and to perform the related action. We call associative this kind of expectations.

In order to explain the planning by imagination mechanism, let us suppose that the robot has perceived the current situation **p**, e.g., it is in a certain position of a room. Let us suppose that the robot knows that its goal **g** is to be in a certain position of another room with a certain orientation. A set of expected situations $\{e_1,e_2,\ldots,\}$ is generated by means of the interaction of both the linguistic and the associative modalities described above. Each e_i in this set can be recognized to be the effect of some action a_i in a set of possible actions $\{a_1,a_2,\ldots,\}$, where each action in the set is geometrically compatible with the current situation **p**.

The robot chooses an action a_i according to some criteria; e.g., a_i is the action whose expected effect has the minimum Euclidean distance in CS from the "goal" **g**, or, for example, considering the emotional valuation of the expected effect. Once that the action to be performed has been chosen, the robot can imagine to execute it by simulating its

effects in CS (see Fig. 7) then it may update the situation and restart the mechanism of generation of expectations until the plan is complete and ready to be executed.

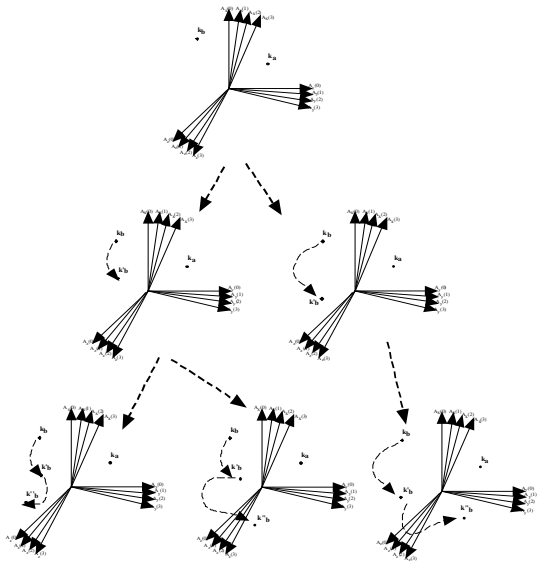


Figure 7: Planning in conceptual space

On the one side, linguistic expectations are the main source of deliberative robot plans: the imagination of the effect of an action is driven by the description of the action in the linguistic KB. This mechanism is similar to the selection of actions in deliberative forward planners. On the other side, associative expectations are at the basis of a more reactive form of planning: in this latter case, perceived situations can "reactively" recall some expected effect of an action.

The linguistic modality of planning may be considered a sort of conscious planning, in the sense that the robot is aware of its course of inferences by means of the symbolic terms involved; these terms, as previously stated, are anchored to c-knoxels.

The associative modality may be considered a form of unconscious planning, as the robot is aware of the results of the associative inferences, but not of the processes that generate them; in facts in this case inferential processes can be considered as some sort of "black box" that associates a set of c-knoxels to another set of c-knoxels, as in the Conscious-Unconscious-Conscious triad mechanism described by Baars (1988).

Both modalities contribute to the full plan that is imagined by the robot when it simulates the plan in the CS. When the robot becomes fully aware of the plan and of its actions, it can generate judgements about its actions and, if necessary, imagine alternative possibilities.

4 The robot at work

Fig. 8 shows our robot in the working environment (left) and the 3D graphical representation of the c-knoxels (right), along with colour and texture. Moreover, the CS of the tour robot will contains many a-priori information as, for example, the map of the robot environment (Fig. 9).

Figure 8: The robot (left) and the 3D view of the robot CS (right).

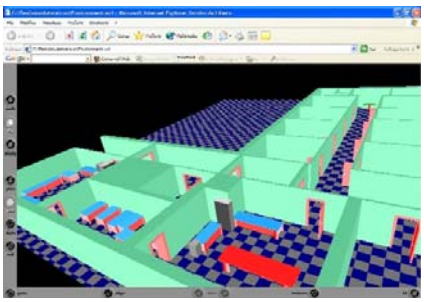


Figure 9: The 3D representation of the robot environment.

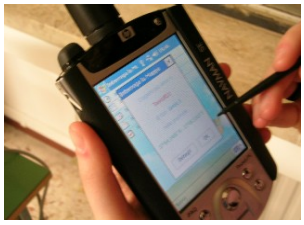


Figure 10: The hand-held computer with GPS.

Let us now consider an example concerning our robot in the Cicerobot framework. The user interfaces with the robot by a hand-held computer equipped with a GPS system (Fig. 10). In the described experiment, the user asks the robot to find a photo of a specific object in the museum and then to go to the user. The linguistic area receives the orders in terms of queries in the OpenCyc symbolic KB (Fig. 11). The hand-held computer also sends the world coordinates of the user by the GPS system.

Figure 11: A picture of the KB of the robot implemented by OpenCyc.

The robot now generates the plan of the sequence of its actions in order to satisfy the user query.

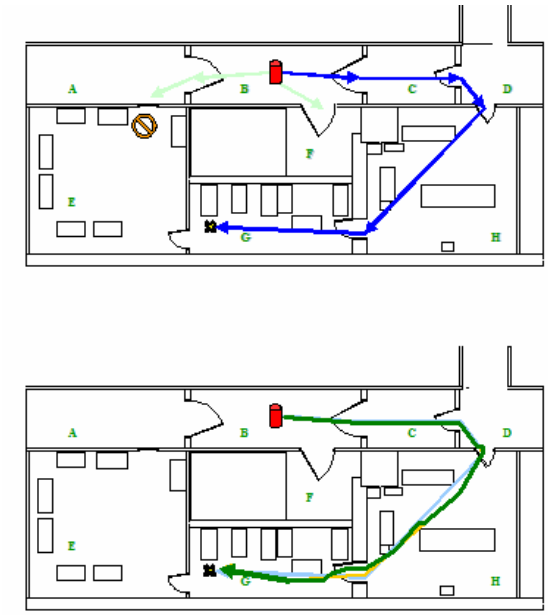


Figure 12: Planning in Conceptual Space

Fig. 12 shows an example of planning in CS. The robot is in the middle of room B and it has to go in room G in the position indicated with the yellow cross. It starts to generate a purely symbolic plan in order to reach the goal (Fig. 12 upper) based on rough symbolic information about the environment. After the plan is complete, the robot imagines itself doing the planned actions in the CS, i.e., it becomes aware of the plan and it is able to judge the plan and analyse the results.

When the robot judges an action as not satisfactory, it imagines possible alternative actions compatible with the current Situation. This process of alternating unconscious planning and conscious imaginations terminates when the sequence of actions is considered as satisfactory (Fig. 12 lower) and the robot is ready to execute the plan.

An example of not satisfactory Action is shown in Fig. 13, where the robot find an unexpected obstacle (the edge of the table) in its imagined trajectory towards the door in the lower left of the figure. When the robot recognizes an instance of the `Blocked_path` situation, it generates the expectations for a `Free_path` situation as the effect of an Avoid action. In this case, the robot generates expectations for a free path in order to find an escape point.

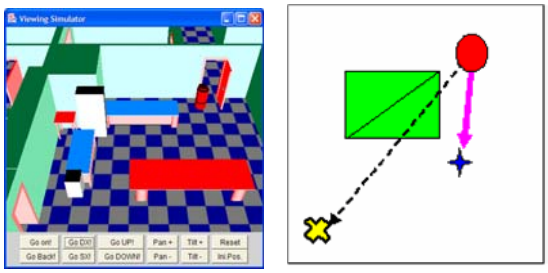


Figure 13: The robot find an unexpected obstacle (the edge of the table) in its trajectory.

During planning, the linguistic area generates all the assertions describing the performed operations, i.e., that the robot has received an order, that it generates a plan to satisfy the order, that it imagines to execute the plan, that it encounter a difficulty and refine its plan, and so on. Fig. 14 shows the console of the robot during the alternation of unconscious planning and conscious action imaginations and the generation of the corresponding assertions.

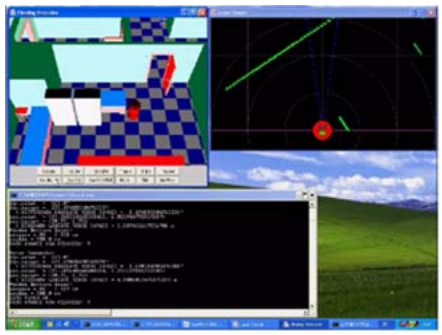


Figure 14: The system console showing the robot simulations during planning.

After this step, the robot is ready to execute the plan. Fig. 15 shows the initial Situation.

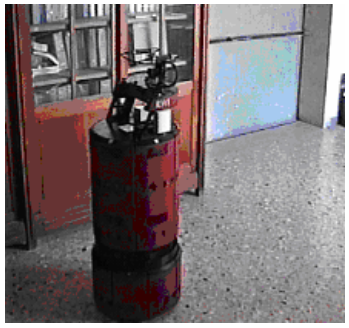


Figure 15: Initial Situation

Fig. 16 shows an "avoid" Action performed by the robot. In the figure, the robot encounter an unexpected person (left), then it stops starting to turn right to avoid the person. As the person continues his motion, the robot may stop to turn and continue its tour (right).

Figure 16: An avoid Action performed by the robot.

Figure 17: A plan transition.

Fig. 17 shows a Situation where the robot runs towards a goal location. Now, the location is currently occupied by other people; the robot imagines itself repeatedly trying to enter in the occupied location. In this case , the robot generates a plan transition in order to supersede the current Situation and to generate a plan patch, for example, to reach another position with another copy of the required object. When the robot terminates its tasks, it will go to the user, as requested. (Fig. 18).

Figure 18: The robot terminates its plan.

Now, the robot has terminated its plan tasks, and it may introspectively remember the whole sequences of actions performed, by the higher-order c-knoxels generated during its operation. The robot may then acquire new skills: e.g., it may learn the successful actions in its memory, in order to be able to associatively recall them at a second time. Moreover, the robot may also learn the unsuccessful and even the dangerous Situations actions, in order to avoid them in the future mission tasks.

The Cicerobot has been experimented at the Archaelogical Museum of Agrigento. The robot in the Museum environment reported the same behaviours previously described. Fig. 19 shows the map of the "Telamone" Hall where the robot operated. Fig. 20 shows images acquired from the robot camera. Fig. 21 shows the 3D robot inner representation of the museum hall.

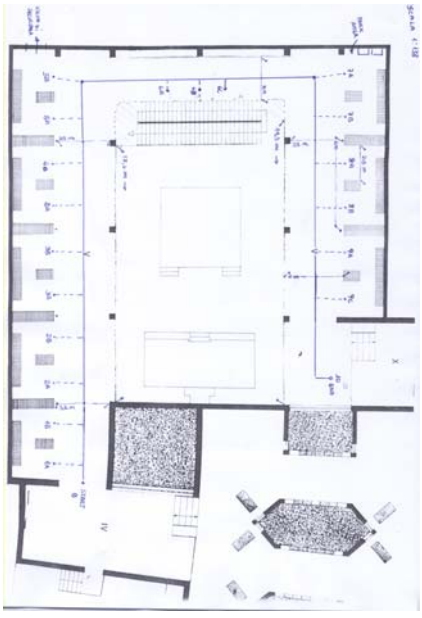


Figure 19: Map of the "Telamone" Hall of the Archaeological Museum of Agrigento.

Figure 20: Images acquired by the robot camera during tours at the Archaeological Museum.

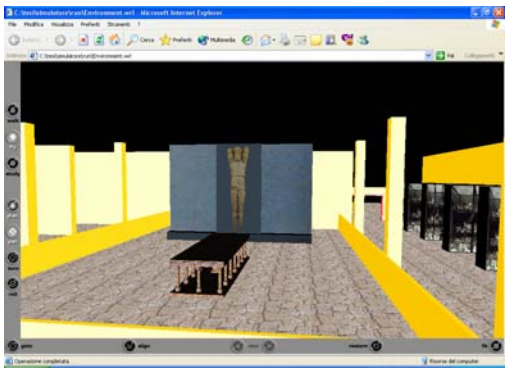


Figure 21: The 3D robot inner representation of the "Telamone" museum hall.

5 Conclusions

In this paper we have presented a cognitive robot architecture based on the integration between subconceptual and linguistic computations through the introduction of the intermediate conceptual space. The architecture is organized in three computational areas. The subconceptual area is concerned with the processing of data coming from the robot sensors. In the linguistic area representation and processing are based on a semantic network formalism. This area is essentially the long-term memory of the robot. The conceptual area is intermediate between the subconceptual and the linguistic areas. Here, data is organized in geometric and "gestaltic" structures in terms of conceptual spaces (Gärdenfors 2000).

The paper also outlined the reciprocal roles of subconceptual computations, conceptual area representations and linguistic knowledge for control of attentive processes, for behaviour planning for affective evaluations and sense of self.

The architecture has been tested on a RWI B21 autonomous robot system on tasks related with guided tours in museum environment. We claim that the proposed architecture addresses the main capacities which are generally addressed by a conscious agent (Aleksander and Dunmann 2003): the capability of representing itself and the external world, of imagining possible evolutions of itself and the world, of paying attentions to the relevant inner and outer events, of planning future actions

The described model of robot consciousness highlights several open problems from the point of view of the computational requirements. First of all, the described architecture requires that the 3D reconstruction of the dynamic scenes perceived by the robot during its tasks should be computed in real time and also the corresponding 2D rendering. At the current state of the art in computer vision and computer graphics literature, this requirement may be satisfied only in case of simple scenes with a few objects where all the motions are slow.

Moreover, the generation of the flow of consciousness requires that the robot should store in the conceptual space at time t all the information of the conceptual spaces at previous times, starting from the beginning of the robot life. This is a hard requirement to be satisfied because of the physical limitations of the robot memory. Some mechanism that lets the robot to summarize its own past experiences should be investigated.

However, we maintain that our proposed architecture is a good starting point to investigate robot consciousness. An interesting point, in the line of (Nagel 1974), is that a robot has a different awareness of the world that we humans may have, because it may be equipped with several perceptive and proprioceptive sensors which have no correspondences in human sensors, like for example the laser rangefinder, the odometer, the GPS, the WiFi or other radio links, and so on.

Therefore, the line of investigation may lead to study new modes of consciousness which may be alternative to human consciousness, as for example the consciousness of an intelligent environment, the consciousness distributed in a network where the robots are network nodes, the consciousness of a multirobot team, the robot with multiple parallel consciousness, and similar kinds of artificial consciousness.

Acknowledgements

Authors would like to thank Igor Aleksander, Maurizio Cardaci, Peter Gärdenfors and Giuseppe Trautteur for helpful discussions about the topics of this paper. Massimo Cossentino, Ignazio Infantino, Irene Macaluso and several students contributed to the implementation of Cicerobot.

References

Igor Aleksander. *Impossible Minds: My Neurons, My Consciousness.* Imperial College Press, London, 1996.

Igor Aleksander and Barry Dunmall. An extension to the hypothesis of the asynchrony of visual consciousness. *Proceedings of the Royal Society B,* 267:197–200, 2000.

Igor Aleksander and Barry Dunmall. Axioms and Tests for the Presence of Minimal Consciousness in Agents. Jou*rnal of Consciousness Studies*, 10, (4–5):7–18, 2003.

Bernard J. Baars. *A Cognitive Theory of Consciousness.* Cambridge University Press, Cambridge, MA, 1988.

Antonio Chella, Marcello Frixione and Salvatore Gaglio. A cognitive architecture for artificial vision. *Artificial Intelligence,* 89:73–111, 1997.

Antonio Chella, Marcello Frixione and Salvatore Gaglio. Understanding dynamic scenes. *Artificial Intelligence*, 123:89-132, 2000.

Stanislas Dehaene and Lionel Naccache. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79, 1–37, 2001.

Daniel Dennett. *Consciousness Explained.* Penguin, London, 1993.

Richard Fikes and Nils Nilsson. STRIPS: a new approach to the application of theorem proving to problem solving. *Artificial Intelligence,* 2:189-208, 1971.

Peter Gärdenfors. *Conceptual Spaces.* MIT Press, Cambridge, MA, 2000.

Owen Holland (ed.). Special issue on Machine Consciousness, *Journal of Consciousness Studies*, 10(4–5), 2003.

Ray Jackendoff. *Consciousness and the Computational Mind.* MIT Press, Cambridge, MA, 1987.

Ales Jaklič, Ales Leonardis and Franc Solina. *Segmentation and Recovery of Superquadrics.* Kluwer Academic Publishers, Boston, MA, 2000

Douglas B. Lenat. CYC: A large-scale investment in knowledge infrastructure. Com*munications of the ACM*, 38(11):32-38, 1995

David Marr. *Vision*. W.H. Freeman, New York, 1982.

Thomas Nagel. What is it like to be a bat? *Philosophical Review,* 83(4):435–50, 1974.

Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing.* Prentice Hall, Englewood Cliffs, NJ, 1989.

Shimon Ullman. *High-level Vision.* MIT Press, Cambridge, MA, 1996.

Semir Zeki and Andreas Bartels. The Asynchrony of Consciousness, *Proceedings of the Royal Society B*, 265:1583–1585, 1998.

Imagination and Abstraction of Sensorimotor Flow: Towards a Robot Model

John Stening*
*University of Skövde
School of Humanities & Informatics
PO Box 408, SE-54128 Skövde
b00johst@student.his.se

Henrik Jacobsson*
*University of Skövde
School of Humanities & Informatics
PO Box 408, SE-54128 Skövde
henrik.jacobsson@his.se

Tom Ziemke*
*University of Skövde
School of Humanities & Informatics
PO Box 408, SE-54128 Skövde
tom.ziemke@his.se

Abstract

This paper presents some initial steps towards a neuro-robotic model of sensorimotor flow abstraction and imagination. Experiments are presented with a two-level neural network architecture combining unsupervised low-level abstraction from sensory and motor values to simple 'concepts' with a higher-level mechanism for prediction and simulation of sequences of such concepts and their respective durations. The analysis of the experiments illustrates a synthetic phenomenology approach to understanding how the robot sees and categorizes the environment while interacting with it, and how imagines it without actually moving.

1 Introduction

Several authors have argued in recent years that substantial parts of cognition and consciousness, in particular imagination and the experience of an 'inner world', can be explained in terms of *simulations* or *emulations* of sensorimotor interaction with the world that allow an agent to temporarily detach its mental processes from the interaction with the outer world (e.g. Clark and Grush, 1999; Hesslow, 2002; Holland and Goodman, 2003; Grush, 2004). There is plenty of empirical evidence for the general idea from neuroscience and psychology (cf. also Svensson et al., 2004, in press), but not yet any convincing computational models that could provide a robot with more than a trivial simulation-based 'inner world'.

Several research groups have tried to implement simulation of perception in robots in a fairly straightforward fashion through the chaining of *forward models*, i.e. the repeated prediction of future sensory input based on current input and planned/simulated action (e.g. Jirenhed et al., 2001; Ziemke et al., 2002, in press; Hoffmann and Möller, 2004; van Dartel et al., 2004). All of these experiments, however, have addressed simulations only at the lowest level of actual sensory input and motor output, and demonstrations of successful simulations have been limited to relatively simple environments, in particular in our own work (cf. Ziemke et al., in press).

The work presented in this paper has therefore been based on the idea of robots internally simulating their sensorimotor interaction with the outer world at some (low) level of abstraction, e.g. in terms of concepts such as 'corridors' and 'corners' and their approximate temporal durations.

The rest of this paper is structured as follows: Section 2 briefly overviews the background and the previous works that our approach has been based on, and Section 3 describes the experiments and their results. The final section then presents a brief discussion of future research directions and ideas for overcoming the limitations of the model presented here.

2 Background and Previous Work

Hesslow's (1994, 2002) *simulation hypothesis* assumes three mechanisms in order to explain the phenomenon of the 'inner world'. The first assumption is *covert behavior*, i.e. the ability to generate neural motor responses that do not become externally observable bodily actions but only neural activation patterns that remain purely internal. Secondly, the existence of a *sensor reactivation* or *imagery* mechanism is assumed. This allows for internally generated activation of sensory areas in the

* Correspondence should be addressed to the third author.

brain, so as to produce the simulated experience of a stimulus, but without the presence of that external stimulus. Finally, the existence of an *anticipation* mechanism is assumed, i.e. the ability to predict or simulate the sensory consequences of a motor response. Support for each of these assumptions can be found in the neuroscience literature (Hesslow, 2002).

With these three mechanisms in place, it should be possible, to *internally simulate* behavioral sequences as illustrated in Figure 1: (a) A situation S_1 elicits activity s1 in the sensory cortex, which in turn leads to a motor response preparation r1. The response preparation r1 results in the overt behavior R_1 which courses a new situation S_2. (b) A predictable relation between a response and the resulting stimuli allows associations to be formed such that the response preparation r1 directly elicits the activity s2 in the sensory cortex. (c) If internally generated stimuli can elicit a response preparation, it should be possible to simulate long sequences of responses and sensory consequences.

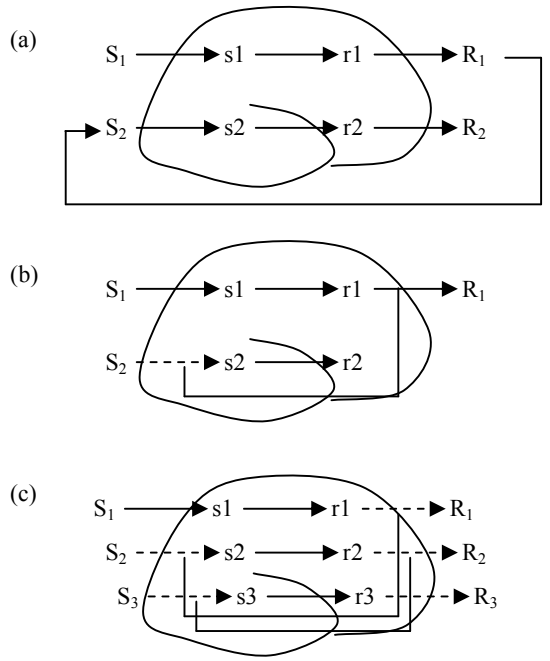


Figure 1: The basic principle of Hesslow's (2002) simulation hypothesis. See text for details.

Several modelers have translated the ideas illustrated in Figure 1 fairly directly into neural robot control architectures that map not only sensory to motor output but also predict the next time step's sensory input, as illustrated in Figure 2.

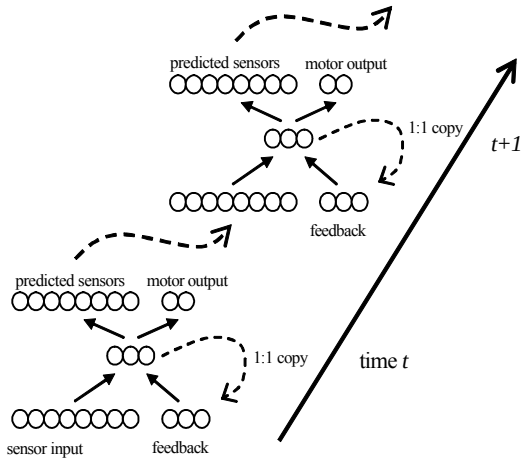


Figure 2: The basic approach to simulation of perception in robots used by Jirenhed et al. (2001) and by others in a similar form. See text for details.

As illustrated in Figure 2, in the experiments of Jirenhed et al. (2001) a neural network trained to map sensory input to motor output was also trained to predict the next sensory input (cf. Ziemke et al., 2002, in press). During simulation then predicted sensor activation was used instead of real sensor input in each time step. Later experiments by several groups have used similar approaches based on a chaining of *forward models*, i.e. the repeated prediction of future sensory input based on current input and planned/simulated action (e.g. Ziemke et al., 2002, in press; Hoffmann and Möller, 2004; van Dartel et al., 2004). All of these experiments, however, have addressed, with varying success, simulations only at the lowest level of actual sensory inputs and motor outputs. Demonstrations of successful simulations have been limited to fairly simple environments, e.g. a square environment with four identical corridors in our own work (cf. Ziemke et al., 2002, in press).

Several theorists have recently also pointed out that a weak spot in simulation/emulation theories is the question of the *level of granularity or abstraction* at which internal simulations might be carried out (cf. Hesslow et al., 2002; Meltzoff and Prinz, 2002; Grush, 2004; Shanahan, subm.; Svensson et al., 2004, in press). The experiments presented in this paper have therefore been based on the idea of robots internally simulating their sensorimotor interaction with the outer world at some low level of abstraction, i.e. in terms of concepts such as 'corridors' and 'corners' and their approximate temporal durations.

3 Experiments

3.1 Architecture

The architecture used in the experiments presented here has been strongly inspired by the two-level architecture used in the work of Nolfi and Tani (1999). In their experiments, the lower level consisted of an unsupervised vector quantizer that categorized current sensory and motor values into more abstract 'concepts', such as 'corner' or 'corridor'. The higher level consisted of a recurrent neural network that predicted the sequence of lower-level concepts and their respective durations (for example, that a corridor lasting twenty time steps would be followed by a left-turning corner of five time steps, etc.).

Our architecture differs from that of Nolfi and Tani (1999) in two respects: Firstly, a different unsupervised vector quantizer, the Adaptive Resource Allocating Vector Quantizer (ARAVQ) of Linåker and Niklasson (2000a, 2000b) was used, which (a) allows for a dynamic number of 'concepts' (model vectors), and (b) is based on the principle of change detection rather than traditional error minimization, which Linåker and Niklasson (2000a) have argued to be better suited for sensory flow segmentation. Furthermore, Linåker and Niklasson (2000b) also presented a mechanism for inverting abstracted concepts 'back' to sensory and motor values. This allows for illustration/imagination in concrete sensorimotor terms, which seems more appropriate for a robotic 'inner world'. For a detailed account of the ARAVQ and its use in the experiments described here see Linåker and Niklasson (2000a, 2000b) and Stening (2004) respectively.

The second main difference to the architecture of Nolfi and Tani (1999) is that in their work the higher-level was only used for prediction of concepts always only one time step ahead, whereas in our case it is used very similar to the networks illustrated in Figure 2. That means, by repeatedly feeding the prediction network's output back to its own input units, the robot is given the capacity to simulate sequences of concepts/categorizations and their respective duration. Our overall architecture consisting of low-level abstraction and higher-level prediction/simulation is illustrated in Figure 3.

It should be noted that, since each of these concepts can be inverted back to its approximate sensory and motor equivalents, we can, at least in principle, let the robot simulate behavioral sequences and illustrate them for our own, the observers', purposes (a case of *"synthetic phenomenology"*). This will be demonstrated later in the analysis of some of the experimental results.

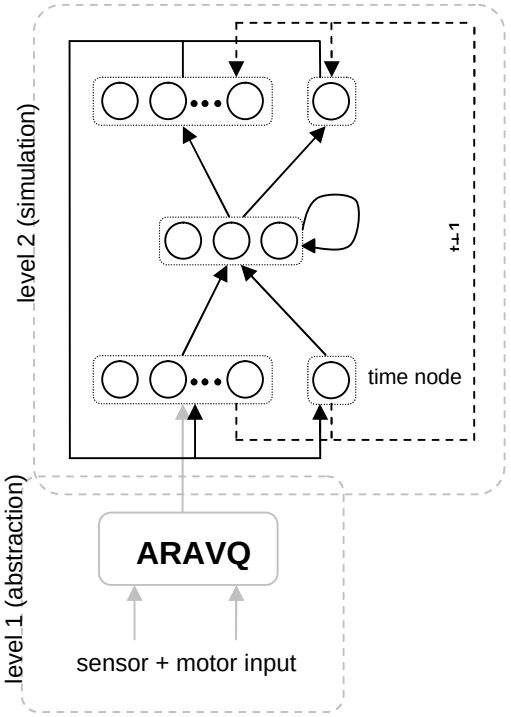


Figure 3: A two-level neural network architecture for sensorimotor flow abstraction and simulation.

3.2 Robot and Environment

The robot is a simulated version (Carlsson and Ziemke, 2001) of the standard Khepera robot (Figure 4). The environments are very similar to those used by Nolfi and Tani (1999), consisting of two rooms connected by a short corridor (Figure 5).

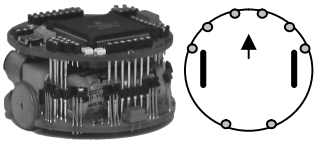


Figure 4: Standard Khepera robot (kteam.com) with eight infrared sensors and two wheels/motors.

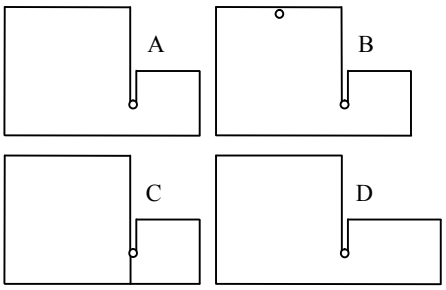


Figure 5: Robot environment (A) and three variations with an extra object (B), a closed door (C) and a longer small room (D).

The robot's behavior is controlled by a pre-trained neural network (not explained here in detail) that generates simple right-hand following behaviour as shown in Figure 6 (details in Stening, 2004).

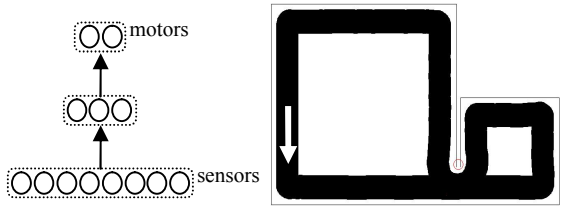


Figure 6: Robot controller and behavior.

3.3 Abstraction

The low-level unsupervised abstraction was tested with different ARAVQ parameters (for details see Stening, 2004). Figures 7 and 8 illustrate the results of unsupervised categorizations of environment A into three and five concepts respectively.

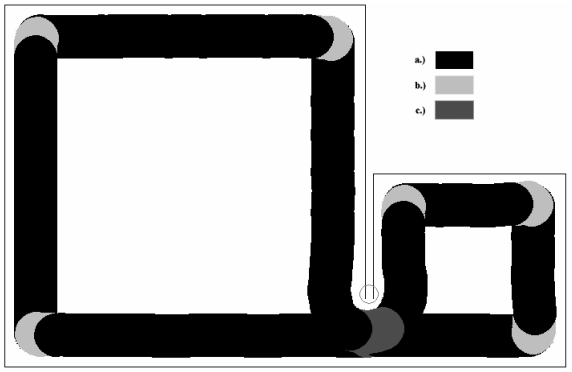


Figure 7: Categorisation (per time step) into three sensorimotor concepts: (a) following a wall to the right (black), (b) turning left (light grey), and (c) moving straight ahead in the corridor (dark grey).

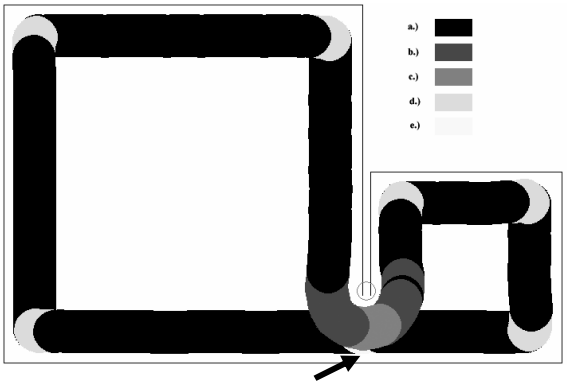


Figure 8: Categorisation into five sensorimotor concepts: (a) following a wall to the right, (b) turning right, (c) turning right in the corridor, (d) turning left in corners, and (e) moving straight in the corridor.

3.4 Prediction

As in the experiments of Nolfi and Tani (1999), the higher-level prediction network was trained on the sequence of concepts and their durations, i.e. to predict the next concept and when it would occur (i.e. the duration of the current segment/concept). The input and output representation for the concepts was a localistic one-bit, winner-take-all representation and the duration (in seconds) was represented by one unit, linearly increased by 0.1 per time step (each time step lasted 0.1 second, e.g. an activation value of 3.7 (seconds) corresponds to 37 time steps).

In the five-concepts case illustrated in Figure 8 the sequence of concepts turned out to be difficult for the network to learn. Neither backpropagation training, as successfully used by Nolfi and Tani (1999) for three and four concepts, nor training with an evolutionary algorithm (for details see Stening, 2004) succeeded in the sense that the sequence of concepts for a whole lap in the environment could be predicted 100% correctly. This is probably due to the fact that the whole-lap sequence is relatively difficult, with 20 transitions from one concept to another and two concepts, (c) and (e), occurring only once. Evolutionary training worked better than backpropagation, but predicted at best 18 out of 20 transitions correctly, failing for concepts (c) and (e).

In the three-concepts case illustrated in Figure 7, backpropagation training achieved at best 15 out of 16 correct predictions in a one-lap sequence, although a large range of training parameters were used, including those used by Nolfi and Tani (1999). Training the prediction network with an evolutionary algorithm (for details see Stening, 2004), on the other hand, was successful in the sense that the sequence of concepts (binary values) could be predicted 100% correct, although the exact durations (real values), naturally, turned out to be much more difficult to predict.

Table 1 shows the real/correct and the predicted (one time step ahead) concepts and their respective durations for two laps (16 concepts each). Note that time steps 17-32 have the same real concepts and durations as time steps 1-16, but the predictions are not identical during the first and the second round. The table shows that all concepts are predicted correctly. Moreover, the durations are predicted roughly correctly for concepts for corners (b) and corridors (c). For wall-following (a), on the other hand, the predicted durations are much more inaccurate, probably due to the fact that they vary much more and the nonlinear nature of the output unit activation function makes it difficult to get the values exactly right (for details see Stening, 2004).

Table 1: Real vs. predicted concepts and durations for the three-concepts case (two laps)

time *t*	real	predicted
1	b 0.6	b 0.5965
2	a 6.1	a 3.9221
3	b 0.5	b 0.6045
4	a 6.1	a 1.5040
5	c 0.2	c 0.2055
6	a 3.5	a 8.4159
7	b 0.6	b 0.5968
8	a 2.5	a 5.5843
9	b 0.6	b 0.5970
10	a 2.5	a 2.5737
11	b 0.6	b 0.5996
12	a 2.9	a 0.4739
13	c 0.8	c 0.0563
14	a 6.7	a 8.8655
15	b 0.6	b 0.5986
16	a 6.0	a 6.0963
17	b 0.6	b 0.6013
18	a 6.1	a 3.6309
19	b 0.5	b 0.6050
20	a 6.5	a 1.2756
21	c 0.1	c 0.2221
22	a 3.4	a 8.3370
23	b 0.6	b 0.5967
24	a 2.5	a 5.5210
25	b 0.6	b 0.5971
26	a 2.6	a 2.5095
27	b 0.6	b 0.6000
28	a 2.7	a 0.4559
29	c 0.6	c 0.0533
30	a 6.8	a 8.8648
31	b 0.6	b 0.5986
32	a 6.1	a 6.1049

Figure 9 provides a graphical comparison between the correct and the predicted sequences of concepts (one time step ahead). This illustrates that the concepts are predicted correctly, but the durations of the segments are much more accurate for corners (b, white) and corridors (c, black) than for wall-following (a, grey).

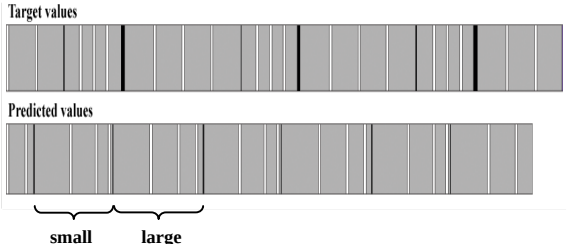


Figure 9: Real target values vs. predictions of concepts (represented by grey level) and durations (represented by width): corners/left turns (b, white), corridors (c, black) and wall-following (grey). NB: The colours are *not* the same as in Figure 7.

3.5 Dealing with change

To test if their prediction network had captured the topology of the environment, Nolfi and Tani (1999) carried out experiments where, for trained robots, they either changed the environment (cf. Figure 5) or the robot's position in it. Very similar experiments were carried out here, but they are reported only briefly in this paper (for details see Stening, 2004).

Experiments on *change detection* were carried out by taking a robot trained in the environment A and testing its prediction performance in the modified environments B (where a round object was added), C (where the corridor was closed, and D (where the small room was longer than before). During the test phase the robot moved in the modified environments and, with no further learning, predicted for a time corresponding to 10,000 concept transitions. The results showed that the robot predicted equally well in D as in A: more than 99% of the concepts were predicted correctly and durations were predicted approximately equally well in A and D, which means that the robot could not really 'tell' the difference between these two environments. In environment B, on the other hand, about 15% of the concepts were predicted incorrectly, due to the added object and duration prediction was worse than in A and D. In environment C, finally, where the corridor had been closed, the prediction did no longer work at all: practically all concept transitions were predicted incorrectly, and duration predictions were much worse than for the other environments (for details see Stening, 2004). That means, as in the experiments of Nolfi and Tani (1999), the prediction network has captured some of the topology of the environment it was trained in (A), and thus can be used to detect changes to it (at least of type B and C).

Experiments on (re-) *localisation* were carried out by taking the robot out of its environment while it was moving and putting it back in at a different position (on the regular trajectory though). As in the experiments of Nolfi and Tani (1999), robots naturally were 'confused' for a couple of time steps and incorrectly predicted one or two concept transitions, but they re-covered very quickly and soon started to predict correctly again (details in Stening, 2004). Again, this shows that the prediction mechanism has captured some of the structure of the environment.

3.6 Internal simulation

The results in the previous subsections concerned the case where the robot moves in the environment and predicts, always only one time step ahead, the next concept and when it will occur. A more difficult test case (and with respect to the discussion in the introduction a more interesting one) is a situation where the robot stands still and internally simulates a whole chain of sensorimotor interactions with the environment, i.e. repeatedly using its own predictions as input to the higher-level network. We tested this in the setup illustrated in Figure 10. The trained robot first moves around in the environment

to localize itself (cf. previous subsection). After ten concept transitions it stops moving (overtly) and internally simulates the sequence of concepts and segment durations as if it were still moving.

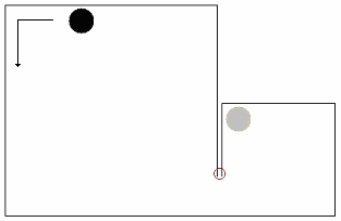


Figure 10: The black circle indicates where the robot starts moving and orienting, the grey one where it stops moving and starts its internal simulation.

As Table 2 and Figure 11 illustrate, the robot's internal simulation is successful roughly to the same degree as in the prediction case: the sequence of concepts is simulated 100% correct, the simulated durations are roughly correct for corners, somewhat underestimated for corridors, and most difficult to simulate for wall following segments. This is here only illustrated for the rest of the first lap (cf. Figure 10) and a complete second lap, but the robot can continue to simulate several laps in a very similar manner (for details see Stening, 2004).

Table 2: Real vs. simulated concepts and durations for the three-concepts case

time *t*	real	Simulated
1	b 0.6	
2	a 6.1	
3	b 0.5	
4	a 6.1	
5	c 0.2	
6	a 3.5	
7	b 0.6	
8	a 2.5	
9	b 0.6	
10	a 2.5	
11	b 0.6	b 0.6020
12	a 2.9	a 1.2294
13	c 0.8	c 0.0285
14	a 6.7	a 8.9167
15	b 0.6	b 0.5991
16	a 6.0	a 6.2807
17	b 0.6	b 0.6012
18	a 6.1	a 3.8592
19	b 0.5	b 0.6020
20	a 6.5	a 1.2293
21	c 0.1	c 0.0285
22	a 3.4	a 8.9167
23	b 0.6	b 0.5991
24	a 2.5	a 6.2807
25	b 0.6	b 0.6012
26	a 2.6	a 3.8592
27	b 0.6	b 0.6020
28	a 2.7	a 1.2293
29	c 0.6	c 0.0285
30	a 6.8	a 8.9167
31	b 0.6	b 0.5991
32	a 6.1	a 6.2807

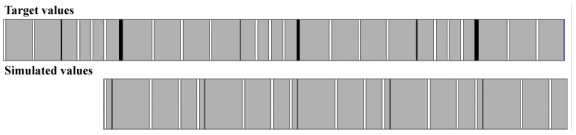


Figure 11: Real target values vs. internal simulation of the sequence of concepts (grey levels) and their durations (widths): corners/left turns (b, white), corridors (c, black) and wall-following (grey).

3.7 Inversion and imagery

The results discussed in the previous subsections illustrate that the robot can, at least in the three-concepts case, more or less correctly predict and simulate the sequence of abstracts concepts and, with some limitations, their durations. From the perspective of a "synthetic phenomenology", however, it is also important to investigate to what degree the robot might be able to, in some sense, 'imagine' the environment, and its own interaction with it, by generating some sort of sensorimotor imagery. Here it should be noted that this should not necessarily be based on a comparison of what the robot 'thinks' the environment looks like and what it 'really' looks like to us as observers, but rather on a comparison of the how the robot 'sees' the world while moving in it and how it simulates or imagines it internally.

As mentioned previously, Linåker and Niklasson (2000b) presented a technique for inverting the concepts (model vectors) abstracted by the ARAVQ, which we used for the low-level quantization of the sensorimotor flow, 'back' to actual sensory and motor values (for details see Stening, 2004). Naturally, the inversion of such an abstraction cannot be 'lossless', since each abstract concepts 'represents' a certain range of sensory and motor activations. Hence, the accuracy of the inversion, i.e. how well it matches the original environment, strongly depends on how well the limited number of abstracted concepts captures the actual range of sensorimotor situations arising in the interaction of agent and environment. Here it will be useful to compare the three-concepts case to the five-concepts case, although we have in the previous subsections focused on the former because the latter could not be predicted/simulated 100% correctly (cf. 3.3).

Figure 12 shows inversions for the three-concepts case (for details see Stening, 2004): (a) an inversion of the real categories/concepts the robot forms while it moves through the environment, (b) an inversion of the predicted values, and (c) and inversion of the simulated values. Start and end point is the upper left corner of the large room.

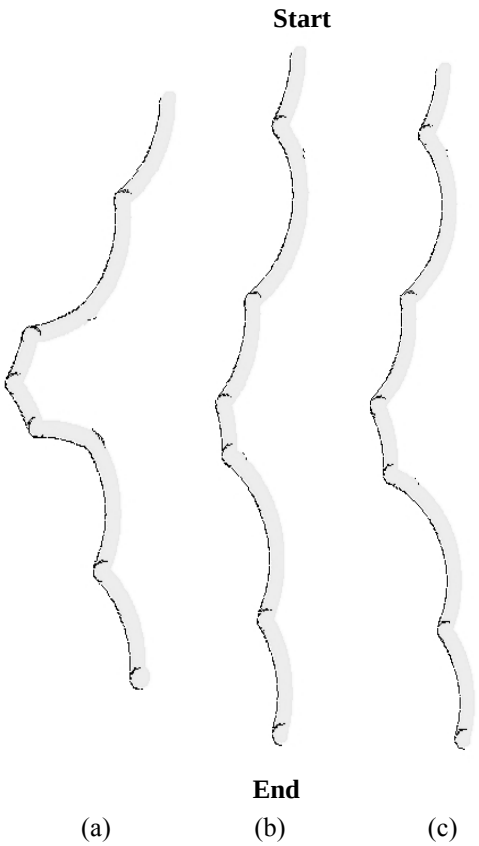


Figure 12: Inversions of sequences of concepts and their durations, based on (a) real values, (b) predicted values, and (c) simulated values. Black dots/lines correspond to sensed walls, and the robot is indicated by one grey circle per time step.

It can be noted that, although the abstract sequence of walls, corners and corridors has been captured correctly, none of the inversions in Figure 12 look particularly similar to our, the observers', view of the environment (cf. Figure 5). This is mostly due to the fact that in the three-concepts categorization (cf. Figure 7) there is no concept corresponding to turning right. Instead the wall-following concept codes for both straight wall-following and most of the right turn before and after the corridor. This distorts the inverted trajectory significantly such that start and end point, which are the same in the original environment (upper left corner of the large room), are not at all close in the inversion. However, as pointed out above, from the perspective of a "synthetic phenomenology" this dissimilarity between the robot's 'imagination' (Figure 12c) and the observers' 'reality' (Figure 5A) is less relevant than the comparison of how the robot 'sees' and categorizes the world while moving in it and how it 'imagines' it while standing still, which is captured *to some degree* by Figures 12a and 12 c respectively, which are much more similar.

The five-concept case, as mentioned above, was analysed in less detail in the work underlying this paper (Stening, 2004), because the sequence of abstract concepts could not be predicted/simulated 100% correctly. Therefore no inversions of predictions and simulations can be presented here. However, Figure 13 illustrates that the inversion of the five-concepts sequence, i.e. how the robot 'sees' and categorizes its world while moving in it, matches much more accurately the structure of the environment as it appears to the observer (cf. Figure 5). To analyze in more detail the robot's predictions and simulations of this sequence might be interesting, although they are not perfect and usually miss the corridor concepts (cf. 3.3 and 3.4).

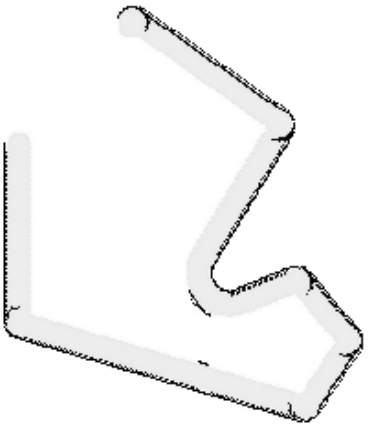


Figure 13: Inversion of the five-concepts sequence of categorizations and durations (real values).

4 Discussion

This paper has presented some initial experiments that aim to contribute towards the development of robot models of sensorimotor abstraction, simulation and imagination. Crucially, unlike most previous experiments with simulation/emulation models, we have here presented a two-level architecture that combines (a) low-level abstraction from sensorimotor values to a limited number of simple abstract 'concepts' (model vectors) with (b) higher-level prediction/simulation of the agent's interaction with the environment at that level of abstraction and (c) the possibility to invert those abstractions back to approximate actual sensory and motor values.

Despite a number of obvious limitations in the current implementation, our experiments show that this architecture successfully learns to simulate the rough structure, i.e. the right sequence of concepts, for environments of the type/complexity Nolfi and Tani (1999) used (two connected rooms of different size), although the duration of segments turns out to

be difficult to estimate. Through inversion of the abstract concepts into prototypical sensory and motor values the sequence of concepts can be retranslated into an inner sensorimotor simulation of the environment that captures the structure of environment, although so far not accurately enough, for example, to allow the robot to navigate the environment completely blindfolded (cf. Ziemke et al., in press).

Some of the crucial points that need to be addressed in future work are the following: Firstly, in the experiments presented here the only connection between action, abstraction, conceptualization, prediction and simulation/imagination is a one-way flow of information. That means, the robots acts completely independent of any needs that it might have, and which might require conceptualization and imagination in the first place. In the current architecture, the abstraction and prediction/simulation mechanisms are mere observers of the robot's behavior. In a more realistic scenario the robot's cognitive processes should certainly serve to effect its behavior, and subsequently its self-maintenance in a presumably dynamic environment where the consequences of actions needs to be anticipated, etc. Secondly, in the current architecture prediction and simulation capacities are limited to the abstract level, whereas simulation at the sensorimotor level has been addressed in other work of ours (e.g. Ziemke et al., in press). In a more realistic scenario, an agent should ideally be able to anticipate/predict/simulate the environment, and its own interaction with it, at multiple levels of abstraction. A useful starting point for this might be a neural network architecture like Tani and Nolfi's (1999) hierarchical mixture of recurrent experts where prediction takes place at different levels.

To conclude, we believe that the work presented here illustrates some promising directions for further experimental investigations of (abstract) imagination of sensorimotor flow, and for further developments of the synthetic phenomenology approach in general.

Acknowledgements

The work described here has profited much from discussions with Germund Hesslow, Dan-Anders Jirenhed, Ron Chrisley, Owen Holland, Murray Shanahan, Nicklas Bergfeldt, Andreas Hansson, Henrik Svensson, and Fredrik Linåker. All experiments were carried out by the first author as part of his masters dissertation project (Stening, 2004).

References

J. Carlsson (Zaxmy) and T. Ziemke. YAKS - Yet Another Khepera Simulator. In: Rückert, Sitte & Witkowski (eds.) *Autonomous Minirobots for Research and Entertainment - Proceedings of the 5th International Heinz Nixdorf Symposium* (pp. 235-241). Paderborn, Germany: HNI-Verlagsschriftenreihe, 2001.

A. Clark and R. Grush. Towards a Cognitive Robotics. *Adaptive Behavior*, 7(1):5-16, 1999.

R. Grush. The emulation theory of representation: motor control, imagery, and perception. *Behavioral and Brain Sciences*, 27(3):377-435, 2004.

G. Hesslow. Will Neuroscience Explain Consciousness?. *Journal of Theoretical Biology*, 171:29-39, 1994

G. Hesslow. Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Science*, 6(6):242-247, 2002.

H. Hoffmann and R. Möller. Action Selection and Mental Transformation Based on a Chain of Forward Models. In S. Schaal, A. Ijspeert, S. Vijayakumar, J. C. T. Hallam, & J. A. Meyer (eds.), *Proceedings of the 8th International Conference on the Simulation of Adaptive Behavior* (S. 213-222). Cambridge, MA: MIT Press, 2004.

O. Holland and R. Goodman. Robots with internal models: a route to machine consciousness? *Journal of Consciousness Studies*, 10(4), 2003.

D.-A. Jirenhed, G. Hesslow and T. Ziemke. Exploring Internal Simulation of Perception in Mobile Robots. In: Arras et al. (eds.) *2001 Fourth European Workshop on Advanced Mobile Robotics - Proceedings* (pp. 107-113). Lund University Cognitive Studies, vol. 86. Lund, Sweden, 2001.

F. Linåker and L. Niklasson. Time series segmentation using an adaptive resource allocating vector quantization network based on change detection. In: *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks*, pp. 323-328, IEEE Computer Society, 2000a.

F. Linåker and L. Niklasson. Extraction and Inversion of Abstract Sensory Flow Representations. In *Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior: From Animals to Animats 6*, pp. 199-208, MIT Press, 2000b.

A. Meltzoff and W. Prinz. An introduction to the imitative mind and brain. In: Meltzoff & Prinz (eds.), *The Imitative Mind: Development, Evolution and Brain Bases,*pp. 1-15. Cambridge, MA: Cambridge University Press, 2002.

S. Nolfi and J. Tani. Extracting regularities in space and time through a cascade of prediction networks: The case of a mobile robot navigating in a structured environment. *Connection Science*, 11(2):129-152, 1999.

M. Shanahan. The Imaginative Mind: Rehearsing Trajectories Through an Abstraction of Sensorimotor Space. Submitted for publication.

J. Stening. Exploring Internal Simulations of Perception in a Mobile Robot using Abstractions. Masters Dissertation HS-IKI-MD-04-009, School of Humanities and Informatics, University of Skövde, Sweden, 2004.

H. Svensson and T. Ziemke Making sense of embodiment: Simulation theories and the sharing of neural circuitry between sensorimotor and cognitive processes. In: K. Forbus, D. Gentner & T. Regier (eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 1309-1314). Mahwah, NJ: Lawrence Erlbaum, 2004

H. Svensson, J. Lindblom and T. Ziemke. In: T. Ziemke, J. Zlatev & R. Frank (eds.), *Body, Language and Mind. Vol. 1: Embodiment*. Mouton de Gruyter, in press.

J. Tani and S. Nolfi. Learning to perceive the world as articulated: An approach for hierarchical learning in sensory-motor systems. *Neural Networks,* 12:1131-1141, 1999.

M. van Dartel, E. Postma and J. van den Herik. Categorisation through internal simulation of perception and behaviour**.** In: L. Schomaker, N. Taatgen & R. Verbrugge (Eds.) *Proceedings of the 16th Belgium-Netherlands Conference on Artificial Intelligence*, Groningen, The Netherlands, 2004.

T. Ziemke, D.-A. Jirenhed and G. Hesslow. Blind Adaptive Behavior Based on Internal Simulation of Perception. Technical Report HS-IDA-TR-02-001, Department of Computer Science, University of Skövde, Sweden, 2002.

T. Ziemke, D.-A. Jirenhed and G. Hesslow. Internal simulation of perception: A minimal neuro-robotic model. *Neurocomputing*, in press.

Thin Phenomenality and Machine Consciousness

Steve Torrance

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

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

Abstract

Current-generation approaches to machine consciousness (MC) have a number of characteristic responses to arguments levelled against the enterprise. These responses tend to marginalize phenomenal consciousness. They do so by presupposing a 'thin' conception of phenomenality which is, in fact, largely shared by anti-computationalist critics of MC. The thin conception sees phenomenality as something that can be easily 'peeled away' from the rest of the physical world. On the thin conception, physiological or neural or functional or organizational features are secondary accompaniments to consciousness rather than primary components of consciousness itself. This inadequate conception bedevils much debate about the nature of consciousness. Can there be a more adequate MC programme, operating on an alternative, 'thick' conception of phenomenality? Recent 'enactive' approaches to consciousness perhaps show some signposts in the right direction.

1 Introduction

In order to prepare the path for next- and future-generation approaches to machine consciousness (MC), I propose to look at some problems in current MC research. Everyone agrees that current MC research has shortcomings – that's why we're here at this workshop. But the ones that I will be pointing out may not be the ones that you think you came here to discuss, or ones you recognize! In pointing out certain inadequacies in current work I do not wish to minimize the value of such work both for pushing forward the frontiers of artificial consciousness, and for understanding the nature of natural consciousness. However work that has great value may nevertheless be subject to unrealistic expectations or shaky presuppositions that need to be brought to light to enable fresh directions to be pursued.

Mine is a philosophical exploration. The excuse for philosophers to get involved with practical research in working MC systems is that the broad goals and presuppositions of such practical research constantly need to be made explicit, evaluated and re-shaped, in the light of the constantly moving horizon of theoretical work in cognitive and consciousness science. The current discussion takes its inspiration from a particular wave in theoretical cognitive science, which has achieved a high profile in the last few years, namely the 'enactive' approach (Varela et al 1991, Thompson 2004).

It's common to hear people who adopt the enactive approach arguing that most previous work in cognitive science has been labouring under various misapprehensions. My aim here is to spell out an argument along those lines, deployed specifically in relation to the field of artificial or machine consciousness.

I shall argue that much existing work in artificial consciousness operates with an inadequate philosophical view of consciousness, which may be called the *thin* (or *shallow*) conception of phenomenality. This conception is in fact also shared by many critics of MC. I will discuss some limitations of thin phenomenality, and then sketch an alterna-

tive conception – *thick* (or *deep*) phenomenality, taking some cues from enactive ways of thinking. I suspect that MC researchers may be rather resistant to the conclusions I come to for they imply that success in achieving machine consciousness may be a lot more remote than is currently thought; and that much of the work to date has been looking for those elusive car-keys under quite the wrong streetlamp. Indeed the right streetlamp may not be on this street or the next, but perhaps in another town or continent!

2 Strong and weak machine consciousness

Echoing Searle (1980), one may distinguish between 'weak' and 'strong' MC. Weak MC seeks to model functional analogues to (or aspects of) consciousness. Strong MC aims to develop computational mechanisms that are *genuinely* conscious, which have consciousness which is as little distinguishable as possible from our own conscious experience. Hanging on the word 'genuine' are, no doubt, a host of begged questions, not to be unduly picked over here. It's common enough to hear people say that such and such a working system is 'genuinely' X - where X is some psychological property - when that system has as little relation to real cases of X-ing as blood oranges[1] have to do with real blood!

Well, it may be easier to get (real) blood out of an orange than out of a stone. And easier to get consciousness out of a machine than out of a stone, perhaps, if the machine is sufficiently elaborately designed? Part of the problem here is that the boundaries of what counts as a machine are intrinsically tentative at any given time, given the continual developments in technology. Turing tried to fix the relevant notion of machinehood in 1950, in a highly restrictive and abstract way. I am not sure how many present-day MC researchers would regard themselves as bound by those restrictions. But unless a clear definition is given of what counts as a machine and what doesn't (for example, are organisms machines, if non-artificial ones?) it's difficult to state clearly what strong MC actually amounts to.

Those who would see themselves as engaged in weak MC will avoid a lot of these kinds of difficulties. They will see the MC enterprise in terms of modelling various aspects of natural consciousness with the purpose of better understanding the latter, rather than duplicating it via a kind of computational trans-substantiation. Those who see their research activity in terms of weak MC goals may nevertheless believe in the realizability of strong MC in principle. What I'm going to say will be relevant to both supporters of strong and of weak MC, but will be particularly relevant to the former.

3 Functional and phenomenal consciousness

A closely associated distinction that may be made is one between 'functional' and 'phenomenal' consciousness. One recent discussion of the distinction between phenomenal and functional consciousness is to be found in Franklin, 2003. The distinction can be taken as a rough-and-ready version of Ned Block's (1995) more carefully worked out, but possibly more specialized, distinction between phenomenal and access consciousness.

Weak MC may be represented as targeting only functional consciousness, while strong MC seeks to target phenomenality as well. That way of putting things may not be thought altogether adequate, however: many supporters of strong MC will deny that there is any sensible distinction between functional and phenomenal consciousness. For those who think the distinction is a valid one, creating a merely functionally conscious mechanism may be seen as a kind of strong MC, in that such a product would instantiate at least one kind of 'genuine' consciousness. Alternatively it might be considered to be kind of midway between weak and strong MC.

Whatever the merits of the notion of merely functional consciousness as opposed to phenomenal consciousness, the idea of phenomenality is often thought not to sit easily within a computational framework. There is a widely shared feeling that computational processes and phenomenal feel are conceptually disjoint categories. The attempt to explain phenomenality in computational terms is regarded by many as a special instance of the 'explanatory gap' (Levine 1983) that is thought to affect any attempt to assimilate consciousness to physicalistic frameworks. Many of those who think the explanatory gap can be bridged in some way or other nevertheless believe that there is an explanatory tension between computation and consciousness. Enthusiasts of MC – particularly strong MC – tend to deal with that tension by reducing, downgrading or avoiding phenomenality in various ways, as we will see.

[1] Or indeed blood-orange flavoured chocolate – a popular brand of chocolate is currently being promoted in that particular flavour!

4 Absent qualia arguments and MC responses

Arguments against the strong MC programme include versions of the absent qualia (AQ) argument. AQ arguments suggest that, for any set of putative computational/functional conditions for phenomenal consciousness, one can always consistently imagine those conditions obtaining but with phenomenal feel absent. To take a classic example, in Ned Block's 'Chinese Nation' argument (Block, 1978), one imagines a scenario meeting our proposed conditions but where the requisite computational operations are performed by some vast population of human operators. Such a scenario may involve much consciousness – all the myriad experiences of the legions of individual participants – but in so doing it leaves no room for the target phenomenal experience supposedly arising out of the computational operations themselves.

AQ-style anti-computationalist arguments in the style of the Chinese Nation describe scenarios where the relevant computational processing is present but where it is very difficult to believe the relevant (or any) conscious states are present. Another kind of AQ argument deals with scenarios where the computational processing is present and where it seems inviting to think that conscious states may be present, but where it is nevertheless insisted a significant doubt may still exist about the existence of such conscious states. Hence, the argument goes, no fully adequate explanatory embedding of phenomenality in computational or cognitive conditions is possible. (For recent versions of AQ-style arguments of that sort see Block, 2002, Prinz 2003).

I will discuss three kinds of MC response to AQ arguments and to general doubts about the computational realizability of consciousness: the eliminativist, the cognitivist, and the agnostic strategies. All these responses, in some way, try to marginalize phenomenality. There may be other strategies, but these are the main ones, as far as I can see.

(a) *The eliminativist strategy:* Supporters of this strategy claim that notions such as phenomenality, qualia, etc., are conceptually confused, scientifically inadequate and unnecessary to the project of artificially creating genuinely conscious beings (Dennett 1991, Harvey 2002, Sloman & Chrisley 2003, Blackmore 2003).

(b) *The cognitivist strategy:* This strategy seeks to reconstrue phenomenal consciousness in terms of cognitive (or cognitive-affective) processes, that are more computationally 'friendly'. Examples are theories that associate consciousness with rich self-modelling processes, or with globally shared information-handling, but there are many other variants. (Baars 1988, Sloman & Chrisley 2003, Holland 2003, etc.)

(c) *The agnostic strategy:* On this strategy it is conceded that perhaps phenomenal consciousness may not be captured within a computational framework, but the claim is made that an important kind of consciousness – e.g. functional consciousness – may be created nonetheless. The question of whether artificial entities which display only this latter kind of consciousness could ever be 'fully' conscious is left open. (Franklin 2003).

These different strategies tend to be combined or to flow into one another. The first two strategies are more easily associated with the strong MC approach, and the third perhaps with the weak MC approach, but this is only a loose principle of grouping.

By associating these various argumentative strategies with a certain conception of consciousness that I wish to criticize, it should not be taken that I think that the authors cited have a *superficial* view of consciousness. On the contrary, all the MC-friendly authors cited offer some very deep insights into aspects of consciousness, both natural and artificial (as the latter might be). However I feel that there is a deep difficulty underlying existing work in the MC area, and this is what I'm trying to bring to light.

5 The 'thin' conception of phenomenal consciousness.

All these strategies rely upon what I call the 'thin' conception of phenomenal consciousness. The thin conception sees phenomenal consciousness rather like the glint on a pair of patent leather shoes. One can imagine someone getting quite philosophically tangled up about how the shine gets to be on the shoe, perhaps taking it to be a rarified, evanescent, extra surface, not equatable with the leather or even with the layer of polish that coats the leather, but which exists rather as a super-layer which somehow sits on top of both. A robust response to such a notion would be to either dismiss the whole idea of the shine as something extra to the shoe or to resort to a 'reductive' physical explanation in terms of the light-reflective properties of particular kinds of surfaces. In a similar way the idea of phenomenal consciousness as something extra to all the information-processing going on in the brain can be either dis-

missed as confusion, or defused by showing how a rich enough information-processing story can capture all the 'specialness' that phenomenality seems to have.

However I would claim that these arguments in defence of strong MC actually buy into a certain view about phenomenal consciousness which is shared by those who reject strong MC. That is, both the anti-computationalist critiques of MC *and* the standard MC responses are based upon a similar, thin conception of phenomenality.

Thus AQ arguments of the sort discussed earlier trade on the apparent ease with which phenomenality can apparently be conceptually peeled away in any imagined scenario where that scenario is described in non-phenomenal terms. A common idea in AQ arguments (particularly 'zombie' variants of such arguments) is that a being can be imagined which has all the outward and internal organizational (i.e. functional) characteristics of a paradigmatically conscious being, but which lacks any 'inner life'. On such a view the phenomenal feel of consciousness is just like the evanescent glint on the patent leather - a special property which obstinately refuses to coalesce with the object's deeper parts. Small wonder, then, that phenomenality may be so easily problematized and emasculated or shelved, as it is within the various MC strategies commonly found.

Such arguments are fed by the idea that 'feel' *is all there is to consciousness* – so that the various physiological or sensorimotor or neural or organizational features investigated by consciousness scientists are secondary accompaniments to the process rather than primary components of the process itself. It is essential to the thin conception, then, that phenomenal feel is conceptually divorcible from any other features in an agent. And being so divorcible, it generates these two opposing philosophical camps, neither of which is able to offer a convincing refutation of the other's position. It is this conceptual detachability, this 'unbearable lightness,' which may be seen as the objectionable feature of the thin conception of phenomenality – the key reason why it leads to the familiar showdown between computationalists and their opponents.

6 Towards an alternative conception of phenomenality

But is there an alternative conception? What might it consist in? What would a 'thicker' or 'deeper' conception of phenomenality consist of? I suggest that it would need to be couched in terms of *essential lived embodiment* – in terms of the real, physical properties of organic, embodied beings who experience conscious subjectivity, plus environmental and intersubjective aspects, as well as in terms of the subjective feeling itself. On an alternative, thick conception, a person's consciousness will be seen, not as conceptually detachable from everything else about that person, but rather as a deeply embedded, multidimensional, embodied, part of that person's nature, whose elements are interleaved in a multiply-stranded complex phenomenon. (See Torrance 2004 for a development of this conception in terms of a 'Grand Inventory' of properties which together make up the 'deep' concept of embodied consciousness.)

On the thick conception, arguments about absent qualia, zombies, and so on, would be harder – perhaps impossible – to state coherently. If phenomenal feel is conceived of as being essentially contextualized in a embodied, living being, then arguments based on supposedly conceivable scenarios where bodily, organic features are all present but the feel is absent will simply lose their force. (Perhaps arguments feeding from such scenarios will not be subject to a knock-down refutation – rather their persuasive force will simply ebb away, as the alternative, essentially embodied, conception of phenomenality is progressively articulated.)

But could there be a strong machine consciousness programme based on a 'thick' conception of phenomenality?. If the 'thick' conception sees phenomenal feel as *deeply embodied,* as conceptually inseparable from the underlying natural organic, living features of biological beings, then what room could there be for the design and development of artificial (non-biological) beings that merited being called 'conscious' in such a sense? Wouldn't the thick conception be taking the MC programme further away from its goal?

I think there are no easy answers to these questions. The thick conception doesn't make the strong MC project any easier – quite the reverse. But it doesn't necessarily make it an unrealizable goal. In building bridges from the human/mammalian consciousness we know to possible artificial forms, our conception of consciousness must necessarily broaden. A Kuhn-style indeterminacy will affect this broadening (the space of discussion isn't, for all that, arbitrary). We shouldn't expect a crisp set of success-conditions for the achievement of 'genuine' (strong) MC. But neither should we expect that such a goal can be ruled out in a peremptory manner by some neat chain of reasoning.

7 Lived embodiment

One source for developing a thick conception of phenomenality is, I suggest, to be found in the enactive approach developed by Varela, Thompson, and Rosch (1991) The enactive approach to mind centres around the idea of 'lived embodiment' mentioned earlier. Such a conception is derived from the writings of Husserl and of Merleau-Ponty, but is also inspired by writings in theoretical biology, particularly work by Maturana and Varela on the so-called autopoietic mode of existence of organisms (see, for example, Maturana and Varela 1987).

The relation between mind, body and organism (or animal existence) has been explored in a recent paper by Robert Hanna and Evan Thompson (2003; see also Thompson 2004, and forthcoming). Hanna and Thompson discuss what they call the 'Mind-body-body problem', which they see as that of reconciling three different ways in which an individual 'I' can be understood. These are:

- as conscious subjectivity (i.e. phenomenality);
- as living, or lived body (*Leib*) with its own perspective or point of view; and
- as a physiological, corporeal, entity investigable within the natural sciences (*Körper*).

How can a single individual incorporate all three of these different natures? Their proposed solution is that the lived embodiment of the individual (Leib) is ontologically basic, and that conscious phenomenality and physical corporeality are two aspects of the lived body. On this account subjectivity is radically embodied, but its embodiment is not that of the merely physical body, but the lived embodiment of organism.

It should be noted that the sense of 'life' which is involved in the notion of 'lived embodiment' is not a purely biological sense (although it relates to the biological sense), but involves selfhood, perspective and purpose. It is a crucial part of the enactive conception of mind and conscious experience, taking its cue from the phenomenology of Husserl and others, that the status of having a mind is intimately related with the process of *living a life* in this autobiographical, rather than just merely biological, sense. Notice how this approach contrasts with traditional approaches to consciousness, as typified by the thin conception. On views of the latter sort consciousness is radically discontinuous with life. In particular (as we have seen), consciousness generates an explanatory gap on such views, and in a way that living doesn't. There is thus claimed to be a logical gulf between experiencing and physical functioning, whereas modern biology has (supposedly) closed any such gulf between being alive and physical functioning. However, on the alternative, enactive, view there is a continuity between phenomenal experience, living one's life as an embodied individual, and having a biological, physical existence. There is no necessity to see a gap more in the one case than in the other.

8 Autopoiesis and MC

There are many other theoretical strands which can be used to explicate the idea of lived embodiment. A central one concerns the idea of what it is to be an autopoietic, or self-recreating, individual. (Varela, 1979, Maturana and Varela, 1987, etc.) An autopoietic system – whether a unicellular or a more complex creature – acts to further its existence within its environment, through the appropriate exchange of its internal components with its surroundings, and via the maintenance of a boundary with its environment. In earlier versions of autopoietic theory, an autopoietic system was a special kind of machine – one which was in continuous activity to maintain its own existence. In recent developments of the notion (Weber & Varela, 2002, Thompson, 2004), autopoiesis is closely tied to the notions of sense-making and teleology: that is, autopoietic self-maintenance is a source or ground of meaning and purpose for that organism (where that meaning or purpose is intrinsic to the organism, rather than something which is merely the product of a pragmatically useful interpretive attribution on the part of an observer). On this view, autopoietic entities are radically different from 'mere' mechanisms, since, unlike the latter, they *enact* their own continued existence, and their own purpose or point of view.

It is a matter of some dispute whether the defining properties of autopoiesis can be found outside the realm of the truly biological, and it is thus an open question as to whether there is any sense in which computationally based constructs could ever be seen as being assimilable to an autopoietic framework – that is as original self-enacting loci of meaning and purpose, or indeed of consciousness. (See, for example, Ruiz-Mirazo and Moreno 2004, McMullin 2004, Bourgine and Stewart 2004.) Clearly, any programme of producing enactive artificial agents would involve a great shift in design philosophy from that which prevails today in most AI or computing science circles. Ezequiel Di Paolo (2003; and forthcoming) is one writer who believes that a programme of developing artificial autopoietic agents, with intrinsic teleology, at least

provides a reasonable research objective. However even he seems to stop short of proclaiming the possibility of computationally-based consciousness, where the latter is understood in this context. Yet in my view, if any MC programme is to succeed in its goal of capturing a conception of consciousness compatible with a fully adequate picture of our own human lived experience, then it has to go down a path of this sort.

9 MC and moral status

This enactively inspired version of the 'thick' conception of consciousness has, I believe, important consequences for how one views the *moral* status of an individual (see Torrance, 2003, 2004). Autopoiesis applies to self-maintaining agents of even the most primitive kind, yet it provides an essential element of what is involved in an adequate conception of highly developed, intelligent autonomous moral agency. Viewing beings as autonomous centres of meaning and purpose, as living and embodied conscious agents that enact their own existence, is, I believe, an important ingredient of building up a moral picture of ourselves, and those we wish to create in our moral image. On this picture, an agent will be seen as an appropriate source of moral agency only because of that agent's status as an self-enacting being that has its own intrinsic purposes, goals and interests. Such beings will be likely to be a source of intrinsic moral concern, as well as, perhaps, an agent endowed with inherent moral responsibilities. They are likely to enter into the web of expectations, obligations and rights that constitutes our social fabric. It is important to this conception of moral agency that MC agents, if they eventualize, will be our companions – participants with us in social existence – rather than just instruments or tools built for scientific exploration or for economic exploitability.

Clearly, the MC quest, when understood in terms of a 'thick', conception of consciousness as lived embodiment, raises important moral questions. One would be guilty of a failure of reflection if one did not see that any genuinely conscious creature that might result from an MC programme informed by such a conception of consciousness, would set us a great deal of moral puzzles – not the least of which is whether such a programme should be even started upon. There is a growing recognition of the inherent moral dimenstions of the MC enterprise. Thomas Metzinger, for example (2003), expounds at some length his view that consciousness in a system is bound up with that system's phenomenal self model (PSM). (I am sure that possessing a PSM in something like Metzinger's sense is a part of what it is to be a 'lived embodiment'; whether it is sufficient remains to be seen. Metzinger writes that the possession of such a PSM will inevitably involve negative as well as positive affective consequences – suffering – for the system, consequences that have a *moral* weight:

> Suffering starts on the level of PSMs. You cannot consciously suffer without having a globally available self-model. The PSM is the decisive neurocomputational instrument not only in developing a host of new cognitive and social skills but also in forcing any strongly conscious system to functionally and representationally appropriate its own disintegration, its own failures and internal conflicts… .The melodrama, but also the potential tragedy of the ego both start on the level of transparent self-modeling. *Therefore we should ban all attempts to create (or even risk the creation of) artificial and postbiotic PSMs from serious academic research.*
> (Metzinger, 2003, 622. My italics)[2]

Metzinger's conclusion may be thought somewhat extreme – but it deserves consideration. The fact that so much discussion of machine consciousness has in the past been conducted more or less in a moral vacuum is itself a testimony to the superficiality of the conception of consciousness that has often operated in the field. Certainly the moral dimensions of entering into an age of artificially conscious creatures need to be very carefully assessed.

10 Conclusion

Machine consciousness research – current and future – has a lot more to do with real consciousness than blood-oranges have to do with real blood. However, the goal of producing a truly conscious machine may be further away than people would like to think To achieve such a goal it is, I am arguing, necessary to radically reprogram one's conception of consciousness, in such a way that consciousness is deeply related to lived embodiment. The resulting revised understanding of machine consciousness will need careful analysis: it is not clear that anything (natural or artificial) that could be conscious in this revised sense could count as a ('mere') machine. At the very least the notion of 'machine' that would need to be operative would have to be very closely intertwined with the notion of 'organism'; artificial consciousness as a field would need to take its inspiration from biology in a

[2] I am grateful to Owen Holland for drawing my attention to this passage from Metzinger's book. See also LeChat 1986, cited by Calverley in his contribution to this symposium.

much more profound sense than is currently envisaged by most in the field.

Also, the considerations proposed here suggest reducing one's confidence in the belief that the strong MC programme might eventually succeed – at least on the basis of the current known technologies. However it cannot be ruled out in principle. Also, it can't be ruled out (as many opponents of MC would do currently) on the basis of arguments which, whether expressly or no, presuppose a 'thin' conception of phenomenality. Nor can arguments to rule it in be successfully launched on the basis of such a conception.

Working out the details of any serious MC programme will involve much further theoretical discussion, which will go hand in hand with actual MC development, but also with an ongoing assessment of how social and moral attitudes towards AI and artificial agents are evolving.

References

Aleksander, I. and Dunmall, B. (2003) Axioms and tests for the presence of minimal consciousness in agents. *Journal of Consciousness Studies*. 10 (4-5), 7-18.

Baars, B. (1988) *A Cognitive Theory of Consciousnes.* Cambridge, England: Cambridge University Press.

Blackmore, S. (2003) Consciousness in meme machines. *Journal of Consciousness Studies*. 10 (4-5), 19-30

Block, N. (1978) Troubles with functionalism. In C.W.Savage, ed., *Minnesota Studies in the Philosophy of Science,* IX, 261-325

Block, N. (1995) On a Confusion about a Function of Consciousness", *Behavioral and Brain Sciences* 18, 2, 227-247

Block, N. (2002) The harder problem of consciousness. *Journal of Philosophy* XCIX, 8, 1-35

Bourgine, P. and Stewart, J. (2004) Autopoiesis and cognition. *Artificial Life* 20 (3) 327-345.

Dennett, D. (1991) *Consciousness Explained*. Boston: Little, Brown.

Di Paolo, E. (2003) Organismically-inspired robotics: Homeostatic adaptation and natural teleology beyond the closed sensorimotor loop. In K. Murase and T. Asakura (Eds.) *Dynamical systems approach to embodiment and sociality*. Adelaide: Advanced Knowledge International, pp.19-42.

Di Paolo, E. (forthcoming) Autopoiesis, adaptivity, teleology, agency. *Phenomenology and the Cognitive Sciences.*

Franklin, S. (2003) IDA: A conscious artefact? *Journal of Consciousness Studies,* 10 (4-5), 47-66

Hanna, R. and Thompson, E. (2003) The mind-body-body problem. *Theoria et Historia Scientiarum: International Journal for Interdisciplinary Studies* 7.

Harvey, I. (2002) Evolving robot consciousness: the easy problems and the rest. In J. Fetzer (ed) *Evolving Consciousness*, Amsterdam: John Benjamins.

Holland, O. and Goodman, R. (2003) Robots with internal models: a route to machine consciousness? *Journal of Consciousness Studies*. 10 (4-5), 77-109

LeChat, M. (1986) Artificial intelligence and ethics: an exercise in moral imagination. *AI Magazine* 7 (2) 70-79.

Levine, J. (1983) Materialism and qualia: the explanatory gap. *Pacific Philosophical Quarterly* 64: 354-361

McMullin, B. (2004) Thirty Years of Computational Autopoiesis: A Review. *Artificial Life* 20 (3). 277-296.

Maturana, H. & Varela, F. (1980) *Autopoiesis and cognition*. Boston: Reidel.

Maturana, H.R. and Varela, F.J. (1987) *The Tree of Knowledge. The Biological Roots of Human Understanding*. Boston: Shambala Press/New Science Library.

Metzinger, T. (2003) *Being No One: The Self-model Theory of Subjectivity.* Cambridge, MA: MIT Press.

Prinz, J. (2003) Level-headed mysterianism and artificial experience. *Journal of Consciousness Studies,* 10 (4-5), 111-132.

Ruiz-Mirazo, K., & Moreno, A. (2004) Basic autonomy as a fundamental step in the synthesis of life. *Artificial Life,* 10 (3). 235 – 260.

Searle, J. (1980) Minds, brains and programs. *The Behavioral and Brain Sciences*. 3. 417-24.

Sloman, A. and Chrisley, R. (2003) Virtual machines and consciousness. *Journal of Consciousness Studies.* 10 (4-5). 133-172.

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

Thompson, E. (forthcoming) Sensorimotor Subjectivity and the Enactive Approach to Experience, *Phenomenology and the Cognitive Sciences.*

Torrance, S. (2003) Artificial Intelligence and Artificial Consciousness: Continuum or Divide? in I. Smit, W. Wallach and G. Lasker (eds), *Cognitive, Emotive And Ethical Aspects Of Decision Making In Humans And In Artificial Intelligence*, *Vol. II*, pp. 25-29, Windsor, Ontario: IIAS.

Torrance, S. (2004) Us and Them: Living with Self-Aware Systems, in I. Smit, W. Wallach and G. Lasker (eds), *Cognitive, Emotive And Ethical Aspects Of Decision Making In Humans And In Artificial Intelligence*, *Vol. III,* Windsor, Ontario: IIAS.

Varela, F. (1979) *Principles of Biological Autonomy*. New York: Elsevier North Holland.

Varela, F., Thompson, E. & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience*. Cambridge, MA: MIT Press, 1991.

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

Consideration of Machine Consciousness in the Context of Mental Therapy from Psychological and Sociological Perspectives

Tatsuya Nomura[⋆†]

[⋆]Department of Media Informatics,
Ryukoku University
1–5, Yokotani, Setaohe–cho, Otsu,
Shiga 520–2194, Japan
nomura@rins.ryukoku.ac.jp

[†]ATR Intelligent Robotics and
Communication Laboratories
2–2, Hikaridai, Seika–cho, Soraku–gun,
Kyoto 619–0288, Japan
nomura@atr.jp

Koichi Takaishi[‡]

[‡] Depatment of Clinical Psychology,
Kyoto Bunkyo University
80 Senzoku, Makishima–cho, Uji,
Kyoto 611–0411, Japan
ko-takaishi@po.kbu.ac.jp

Tatsunori Hashido[§]

[§]Graduate School of Corporate Information,
Hannan University
5–4–33 AmamiHigashi, Matsubara,
Osaka 580–8502, Japan
mc03013@hannan-u.ac.jp

Abstract

Realization of machine consciousness has a lot of scientific and engineering implications such as clarification of organization of human consciousness, implementation of real humanoid robots and virtual human agents, and so on. However, the most important is not how machine consciousness can be realized, but how people feel for robots and software agents when they recognize that the robots and agents have their own consciousness, and how the society is influenced by the result. This paper discusses these problems from psychological and sociological perspectives.

1 Introduction

Realization of machine consciousness has a lot of implications. Scientifically, organization of human consciousness will be clarified through process of the realization. From engineering perspectives, it will lead to implementation of real humanoid robots and virtual human agents. Then, these robots and agents will be popularized in daily-life. When this popularization of machine consciousness is considered, however, the most important is not how machine consciousness can be realized, but how people feel for robots and software agents when they recognize that the robots and agents have their own consciousness, and how the society is influenced by the result.

These problems are critical when machine consciousness is applied to psychiatric fields such as mental therapy. In fact, even current machines not having consciousness such as empathy have been starting to be applied to mental therapy (Shibata, 1999; Hashimoto, 2001; Dautenhahn and Billard, 2002; Fujino, 2003; Dautenhahn et al., 2002). Turkle (1995) reported that artificial agent programs for psychiatry have socially been allowed since 1990's, while synchronized by standardization of diagnosis and treatment in psychiatry. Moreover, a lot of studies on robotic therapy have recently been done, in particular, in Japan (for example, Shibata (1999); Tashima et al. (1999); Hashimoto (2001)).

In fact, humans are affected even by these machines. Reeves and Nass (1996) showed that experiences of human through artificial media including computers are essentially equal to real experiences, by application of theories in social sciences on human interaction and experimental methods in psychology. In other words, humans unconsciously react to machines in the same way as to humans even if it is consciouslly recognized that those whom they interact are just machines. Some important results by Reeves and Nass (1996) are summarized as follows:

- Humans tend to politely react to machines sending polite messages, and prefer machines sending messages of praise to those sending critical messages.

- Humans tend to interpret even image objects consisting of simple lines as ones having characters by using the same dimensions as those for humans (dominant–obedient and cooperative–non cooperative). In addition, dominant persons prefer machines displaying texts in dominant ways and obedient persons prefer machines displaying texts in obedient ways.

- Humans tend to be affected by social roles assigned with machines such as professionals, teammates, and genders.

- Humans tend to firstly feel good or evil emotions for information even from artificial media, and have a bias to negative information on attention and memory in the same way as the real world.

Important is that these reactions of human are unconsciously evoked. As a source of these phenomena, Reeves and Nass (1996) assume that they are a trace of evolution of mental mechanisms in the ancient wild environment.

Moreover, Turkle (1995) investigated minds of people on computers by interviewing with a lot of people in some countries from 1970's to 1980's. Some important results are summarized as follows:

- As mechanisms of computers became more complex, that is, they lost "transparency", users gave up trying to understand computers based on the physical functions.

- Furthermore, as interactivity of computers became increasing, people began to understand these interactive and nontransparent objects by analogy of mental states of humans, in other words, regard them as objects having mental states, which is not either just physical objects or living ones.

- Researches on artificial intelligence and biologically inspired models such as neural networks have positively been affecting this trend. As one of the results, artificial agent programs for psychiatry have socially been allowed since 1990's, while synchronized by standardization of diagnosis and treatment in psychiatry.

In addition, Turkle (1995) mentioned Eliza effect showing that humans tend to overestimate intelligent capability of computer programs.

These statements imply that even current machines can affect mental states of human by using actions based on their characters and social roles, regardless of positive or negative direction. Moreover, they imply that machine consciousness may really be introduced in psychiatric fields in future. However, it has sufficiently not been investigated what influences these machines have to clients in mental therapy. It has not been denied yet that machine consciousness in the therapeutic contexts have some evil effects to the clients dependent on the cultural situations where they live.

This paper discusses what meanings machine consciousness has in the context of mental therapy, that is, what conditions machine consciousness should satisfy when it is applied to mental therapy, whether machine consciousness can have positive effects when these conditions are satisfied, and what happens when it is introduced in the current social situations, from some perspectives of psychology and sociology.

2 Necessary Conditions of Machine Consciousness in Mental Therapy

This section considers conditions that machine consciousness should satisfy when it is applied as a substitution of psychotherapists, from methodological perspectives of clinical psychology. [1].

2.1 Judgment of Timing

First, it is considered to be valid in all methods of psychotherapy that important is timing in treatment. It means that it is necessary to execute appropriate treatment in appropriate time in psychotherapy. In other words, timing in psychotherapy is a key determining whether the treatment succeeds or not.

However, it is just therapists that judge timing. Therapists individually judge it and take appropriate correspondence for clients. This individual judgment needs the therapists' insight to see through conditions in which the clients stay. This judgment is dependent on perspectives of the therapists on the clients and changed by the therapists' methods and experiences in psychotherapy. Thus, even if a client has the symptom same as another client, it may not be guaranteed that a treatment method effective for the latter is also effective for the former.

This judgment of timing in psychotherapy should also appropriately be executed by machine consciousness. In other words, machine consciousness should have its own experiences and construct its own insight

[1] This consideration is based on Hashido and Nomura (2004)

to see through conditions in which humans stay based on the experiences.

2.2 Rapport

Second, the paradigm of clinical psychology implies that construction of a well relation between a therapist and client needs sympathy, warmth, and beliefs of them. This problem leads us to a concept of "Rapport". Rapport is a state of relations that persons feel friendly and can confidentially talk with each other. It is necessary between clients and therapists in psychotherapy and can exist on confidential relationships between them. Mutual confidence in persons requires their mutual understanding.

Thus, machine consciousness should construct confidential relationships with humans based on its emotion of empathy.

2.3 Pragmatic Analysis and Irregularity

Third, human communication is based on pragmatic analysis for sentences, which deals with "indirect" or "deep" meanings of sentences, in contrast with semantic analysis which deals with "direct" or "surface" meaning. In other words, natural language processing in processes of communication requires not only surface processing but also understanding meanings contained in the background, according to the situations. If this process is executed by machine consciousness, much knowledge of humans' social behaviors and mental states are necessary.

Moreover, daily life conversation between humans includes irregularity not following grammar and contexts. In other words, it has rich unexpectedness, which means that it is not predictive what utterances appear. This irregularity may also be a hint of treatment (for example, there is a method in family therapy where this irregularity is explicitly explored). If machine consciousness copes with this unexpectedness, a wide range of exception handling must be performed. It should also be necessary to prevent machine consciousness from executing inappropriate replies for humans' unexpected behaviors. This type of processing is the most important in psychotherapy.

Thus, machine consciousness should execute pragmatic analysis and cope with irregularity based on much knowledge of humans' social behaviors and mental states.

3 Software and Robotic Therapy under "Psychologism"

Even if machine consciousness having functions mentioned in the previous section are realized, it does not mean that it is useful in mental therapy. Then, this consideration needs psychological and sociological perspectives.

The cultural trend called "psychologism" implies possibility of evil effects of machine consciousness to clients in mental therapy. This word refers to a trend in modern society where psychiatric symptoms in individuals are internalized although they may be caused by social structures and cultural customs, and, as a result, the root social and cultural situations that need to be clarified are concealed.

In this section, we refer to sociological criticism for psychologism and consider implications from it [2].

3.1 Criticism for Psychologism from the Sociology of Emotions and Clinical Sociology

Mori (2000) focused on psychologism on discussing the extreme self–control of people in the modern society. His theory is based on the theory of feeling rules by Hochschild (1983) and the theory of McDonaldozation of Society (rationalization) by Ritzer (1996), and is summarized as follows:

- In the modern society we are always forced to pay attention to our and others' emotions in order not to hurt our emotions each other (cult of personality). Moreover, this cult of personality and psychologism has been complementing each other.

- Furthermore, psychologism and rationalization in the modern society has also been complementing each other, and as a result we are required to have a high degree of self–control for our emotions.

- Persons executing a high degree of emotion management cannot permit others' deviation from feeling rules they observe even if it is only a little. This strict observance of feeling rules and difference of the rules between individuals cause disagreement in the modern society (e.g., increasing child abuse in Japan).

[2]This consideration is based on Nomura and Tejima (2002); Nomura (2003b)

In addition, Mori (2000) claimed based on analysis of increasing psychological manuals for self–helping that psychological knowledge strengthens the social trend of self–control for emotions.

Psychologism has also been criticized in the research field of clinical sociology. Ozawa (2000) criticized the trend that people in the modern societies are dependent on counseling due to psychologism and the extreme emotion management.

3.2 Implications for Machine Consciousness

The above statements from the sociology of emotions and clinical sociology imply that people in the modern society are always required to execute emotion management and dependent on mental therapy for it. Moreover, rationalism as Ritzer (1996) pointed out may also encourage reduction of man power in mental therapy. In addition, Turkle (1995) implies that software agents and robots may be introduced in psychiatric fields. As a result, software and robotic therapy with machine consciousness including emotions such as sympathy may be encouraged.

On the other hand, these people are sensitive for others' emotion management and there is difference of feeling rules between individuals. Thus, these people are also sensitive for emotional behaviors in the software and robots and there is possibility that the emotional behaviors of the systems are not suitable for feeling rules of the clients. Furthermore, it is not clear how people feel for empathy from machines when they appear as substitution of human therapists since the people regard the machines as objects having mental states which is not either just physical objects or living ones, while sensitive to the machine consciousness and being unconsciously reacting in interaction with it (Reeves and Nass, 1996). Thus machine consciousness may give the clients mental burden of emotion management in interaction between them and the systems, and influence therapeutic effects even if they are exactly implemented along theories on therapy.

4 Machine Consciousness as a Popular Product in Mental Therapy

The previous sections consider influence of machine consciousness in mental therapy at a level of individuals. This section considers possibility that machine consciousness is supplied as a popular product in mental therapy, from some sociological perspectives [3].

4.1 Perspectives from the Sociology of Health and Illness

There is another sociological research related to psychologism, called "the sociology of health and illness". According to Nomura (2000), the sociology of health and illness is one of reflective sociological research actions that analyze and criticize discourses on health, dominant in the modern societies, and its theory is based on social constructionism (Lupton, 1994). The research subjects in the sociology of health and illness are relations between cultures and ways of using the concept of "health", mutual interaction between medical staffs and clients, powers that the concept of "health" can have in the societies.

As one of sociological researches of health and illness, Ukigaya (2000) analyzed the social situation on life–style related disease, focusing on diabetes. In her research it was clarified that advertisement of the concept of life–style related disease by the government extremely requires individuals' accountability for health, relations between appearance of the illness and the social situations are concealed as a result, and clients of diabetes are socially and mentally pressed under requirement of self–accountability for their health.

Moreover, it was reported that some clients of diabetes develop their original interpretation of medical knowledge on the illness, and distort it. For examples, some clients are not perfectly ruled by medical indication, such as on meals and sports, but have their original meals and sports according to their bodily and mental states.

4.2 Another Perspective from Narrative Therapy

On the other hand, Asano (2001) pointed out that a new type of mental therapy called narrative therapy (McNamee and Gergen, 1992) has a factor to become popular in the modern societies. He claimed as follows:

- The action to talk narratives on selves is one of cultural practices popular in the modern society, that is, there are a lot of increasing people to want to talk narratives on themselves in USA, Europe, and Japan.

[3] This consideration is based on Nomura and Tejima (2002); Nomura (2003b)

- The modern society has a characteristic to produce these people, and industries aiming at satisfying demand of these people like manuals for making narratives on selves, publishers, and so on, called "narrative industries", have appeared. We should note that narrative therapy is just one of these industries.

- Narrative therapy functions by explicitly drawing things concealed in narratives which clients talk on themselves. However, the desire of people to talk narratives on themselves is also a desire to leave these concealed things concealed. If narrative theorists are not conscious for this fact, narrative therapy has a danger that it only repeats this desire of people

4.3 Implications for Machine Consciousness

The statements by Ukigaya (2000) have some important implications. The concept of life–style related disease and psychologism have the common social power in the sense that both of them press people with some symptoms under self–control of individuals and conceals social situations related with sources of the illness. As a result of it, clients of mental therapy, in particular, software and robotic therapy may develop their original interpretation of psychological knowledge that is presupposition in implementation of these systems, and sometimes distort it. Moreover, the statements by Asano (2001) imply that reception of machine consciousness in the modern society and narrative industries satisfying desires of people to talk on themselves are combined with each other, and as a result machine consciousness appear as a product to help people to make narratives on themselves through interaction with people.

These facts imply problems in cases that therapeutic software and robots are supplied as popular products, not via medical organizations. People having their original interpretation of psychological knowledge with distortion may tend to prefer popular systems suitable for their interpretation of psychological knowledge to systems that are scientifically investigated and selected via medical markets.

However, interaction with machine consciousness may just repeat desire of people to talk on themselves while leaving concealed things concealed in their narratives, which should be drawn in narrative therapeutic conversation between clients and therapists. This situation can happen since narrative therapy does not mean a concrete therapeutic technique but just an attitude that therapists should have for clients, and thus it has rooms for clients' original interpretation for it.

5 Double Bind Situations in Mental Therapy by Machine Consciousness

This section deals with the deeper relations between the cultural trends of mental health and the clients' personal traits in mental therapy using software agents and robots, that is, the relations between psychologism and anxiety traits for computers and robots. We then suggest that clients of mental therapy using machine consciousness may be forced into a kind of double bind situation (Bateson, 1972) [4].

5.1 Anxiety for Computers and Robots

The concept of computer anxiety means anxious emotions that prevent humans from using and learning computers in educational situations and daily life (Hirata, 1990; Raub, 1981). Anxiety can generally be classified into two categories: state and trait anxiety. Trait anxiety is a kind of personal characteristics that is stable in individuals. State anxiety can be changed depending on the situation and time, and computer anxiety is classified into this category. From the perspective of education, computer anxiety in individuals should and can be reduced by educationally appropriate programs, and several psychological scales for its measurement have been developed (Hirata, 1990; Raub, 1981).

On the other hand, from the perspective of mental therapy, computer anxiety can influence the therapeutic effect of software therapy using machine consciousness. If the client's computer anxiety is high, it can prevent interaction with the therapeutic software agents even if the agents are designed based on theories of mental therapy. Of course, communication anxiety should be considered even in therapy with human therapists (Pribyl et al., 1998). However, it can be reduced during the therapy process by the therapist's careful treatment. It is not clear whether computer anxiety can be reduced during the therapy processes by software agents. Thus, anxiety should either be reduced before therapy proceeds or another person should assist clients in interacting with the software agents during the therapy process.

Similar emotions of anxiety should also be considered in robotic therapy. Robotic therapy may be different from therapy using software agents in the

[4] This consideration is based on Nomura (2003a,b)

sense that robots have concrete bodies and can influence client's cognition. Thus, anxiety toward robots may be different from computer anxiety. However, it should be considered that anxiety may be caused by highly technological objects and communication with them. In this sense, anxiety toward robots is a complex emotion of computer anxiety and communication anxiety (Nomura and Kanda, 2003; Nomura et al., 2004).

5.2 Double Bind Theory

The Double Bind Theory, proposed as a source of schizophrenia from the viewpoint of social interactions in the 1950s (Bateson, 1972), argues that schizophrenia may result from not only impact on the mental level of individuals, such as trauma, but also inconsistency in human communication. The conditions for double bind are formalized as:

1. the existence of one victim (a child in many cases) and an assailant or some assailants (the mother in many cases,)
2. the customization of cognition for double bind structures through repeated experience,
3. the first prohibition message with punishment,
4. the second prohibition message inconsistent with the first one at another level (inconsistent situations,)
5. a third message that prohibits the victim from stepping out of the inconsistent situation (prohibition of the victim's movement to a meta level of communication.)

It is pointed out that the double bind theory itself has largely not been developed in the theoretical sense since the 1970s (Ciompi, 1982), and there has not been enough empirical evidence showing that double bind situations are a source of schizophrenia (Koopmans, 1997). Even if the double bind situations are not a source of schizophrenia, however, the double bind theory has been applied in the clinical field as a basic concept of family system theory (Foley, 1986), and it is said that double bind situations frequently exist in daily life.

5.3 Implications for Machine Consciousness

The consideration in the previous sections imply that people in modern society are required to execute emotion management and are dependent on mental therapy for it. In addition, modern rationalism may also encourage a reduction of manpower in mental therapy, and, as a result, software and robotic therapy using machine consciousness may be encouraged. Then, people in modern society may be forced to face therapeutic software agents and robots by the social pressure of the self–control of their emotions and mental health, and rationalism, in particular, if mental therapy becomes a duty of members in organizations such as businesses and schools.

If these therapies are introduced without consideration of anxiety that individuals may experience in their interaction with computers and robots, however, they may cause double bind situations for these individuals, of which clients with high anxiety for computers and robots are victims. These clients are forced to face these systems by social pressure, but they cannot get sufficient therapeutic benefit due to their anxiety for the systems if their anxiety is not reduced by appropriate treatment, due to rationalism in the therapy process. Furthermore, social pressure prohibits them from stepping out of these situations because it signifies their rejection of accountability for their own mental health. In other words, this type of client cannot be treated with software or robotic therapy even if these software agents and robots are designed based on theories of mental therapy.

6 Summary

This paper considered problems of machine consciousness in the context of mental therapy from some perspectives of psychology and sociology.

As discussed the above, even current machines without consciousness have a possibility of introduction in psychiatric fields while few sufficient investigation of their therapeutic effects. Even if they have consciousness needed for substitution of human counselors, it is not clear whether they have sufficient therapeutic effects, due to extreme sensitivity to emotions and social pressure of self–control for mental health caused by psychologism. Moreover, social pressure of self–control for mental health may encourage machine consciousness as a popular product of narrative industries in mental therapy, which just repeats desire of people to talk on themselves while leaving concealed things concealed in their narratives. Furthermore, introduction of machine consciousness in mental therapy at organizational levels may cause double–bind situations for clients having anxiety toward computers and robots, due to social pressure of self–control for mental health.

Of course, it should be discussed whether machine

consciousness can really have emotions such as empathy, in the sense that it has the same organization as humans. [5] As far as humans regard software agents and robots as those having their own consciousness, however, its psychological and sociological influences should be considered, regardless that they can really have their own consciousness or not.

Finally, it should be noted that this paper does not aim at denying application of machine consciousness to psychiatric fields. There may be some disciplines familiar to software agents and robots in mental therapy. However, we should be careful of a naive idea that machine consciousness produces familiarity of humans with machines and realization of it leads to healing pains of clients in therapeutic contexts. As far as we consider applications of machine consciousness to therapeutic fields, we should pay our attention to influences of them in mental, social, and cultural levels.

References

T. Asano. *Narrative–Theoretic Approach to Selves*. Keiso–Shobo, 2001. (in Japanese).

G. Bateson. *Steps to an Ecology of Mind*. Harper & Row, 1972. (Japanese translation: Y. Sato. Shisaku–Sha, 1990.).

L. Ciompi. *Affektlogik*. Klett–Cotta, 1982. (Japanese translation: M. Matsumoto et al. (1994). Gakuju Shoin).

K. Dautenhahn and A. Billard. Games children with autism can play with robota, a humanoid robotic doll. In S. Keates, P. J. Clarkson, P. M. Langdon, and P. Robinson, editors, *Universal Access and Assitive Technology*, pages 179–190. Springer–Verlag, 2002.

K. Dautenhahn, A. H. Bond, L. Cañamero, and B. Edmonds. *Socially Intelligent Agents: Creating Relationships with Computers and Robots*. Kluwer Academic Publishers, 2002.

V. D. Foley. *An Introduction to Family Therapy*. Allyn & Bacon, 1986. (Japanese translation: A. Fujinawa, et al. Sogensha, 1993.).

[5] For example, it can be discussed whether machine consciousness can be implemented on state spaces under the assumption that machine consciousness is an autopoietic system (Nomura, 2001, 2002).

H. Fujino. Current situations and problems of software counseling. In *Proc. IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 571–576, 2003.

T. Hashido and T. Nomura. Consideration on mental therapy using computers. In *Proc. Joint 2nd International Conference on Soft Computing and Intelligent Systems and 5th International Symposium on Advanced Intelligent Systems*, 2004.

T. Hashimoto. Emotion model in robot assited activity. In *Proc. 2001 IEEE International Symposium on Computational Intelligence in Robotics and Automation (IEEE CIRA 2001)*, pages 184–188, 2001.

K. Hirata. The concept of computer anxiety and measurement of it. *Bulletin of Aichi University of Education*, 39:203–212, February 1990. (in Japanese).

A. R. Hochschild. *The Managed Heart*. University of California Press, 1983. (Japanese translation: J. Ishikawa and A. Murofushi. Sekaishishosha, 2000.).

M. Koopmans. Schizophrenia and the Family: Double Bind Theory Revisited. *Dynamical Psychology*, 1997. http://goertzel.org/dynapsyc/1997/Koopmans.html (electric journal).

D. Lupton. *Medicine as Culture: Illness, Disease and the Body in Western Societies*. Sage, London, 1994.

S. McNamee and K. J. Gergen. *Therapy as Social Construction*. Sage, 1992. (Japanese translation: Y. Noguchi and N. Nomura. Kongo–Shuppan, 1997.).

S. Mori. *A Cage of Self–Control*. Kodansha, 2000. (in Japanese).

K. Nomura. Introduction of critical theories for health discourses. In *Addicted to Health Discourses*, chapter 7. Bunka Shobo Hakubun–sha, 2000. (in Japanese).

T. Nomura. Formal description of autopoiesis based on the theory of category. In J. Kelemen and P. Sosik, editors, *Advances in Artificial Life: 6th European Conference, ECAL 2001, Proceedings*, pages 700–703, 2001.

T. Nomura. Formal description of autopoiesis for analytic models of life and social systems. In *Proc.*

the 8th International Conference on Artificial Life, pages 15–18,, 2002.

T. Nomura. Double bind situations in man–machine interaction under contexts of mental therapy. In T. Rist, et. al., editor, *Intelligent Virtual Agents: 4th International Workshop, IVA 2003*, pages 67–71, 2003a.

T. Nomura. Problems of artificial emotions in mental therapy. In *Proc. IEEE International Symposium on Computational Intelligence in Robotics and Automation (IEEE CIRA 2003)*, pages 567–570, 2003b.

T. Nomura and T. Kanda. On proposing the concept of robot anxiety and considering measurement of it,. In *Proc. the 12th IEEE International Workshop on Robot and Human Interactive Communication*, pages 373–378, 2003.

T. Nomura, T. Kanda, T. Suzuki, and K. Kato. Psychology in human–robot communication: An attempt through investigation of negative attitudes and anxiety toward robots. In *Proc. the 13th IEEE International Workshop on Robot and Human Interactive Communication*, pages 35–40, 2004.

T. Nomura and N. Tejima. Critical consideration of applications of affective robots to mental therapy from psychological and sociological perspectives. In *Proc. 11th IEEE International Workshop on Robot and Human Interactive Communication (ROMAN 2002)*, pages 99–104, 2002.

M. Ozawa. Can "age of minds" help people? *Gendai Shiso*, 28–9:92–99, 2000. (in Japanese).

C. B. Pribyl, J. A. Keaten, M. Sakamoto, and F. Koshikawa. Assessing the cross–cultural content validity of the Personal Report of Communication Apprehension scale (PRCA–24). *Japanese Psychological Research*, 40:47–53, 1998.

A. C. Raub. *Correlates of comuter anxiety in college students*. PhD thesis, University of Pennsylvania, 1981.

B. Reeves and C. Nass. *Media Equation*. Cambridge Press, 1996. (Japanese translation: H. Hosoma. Shoeisha, 2001.).

G. Ritzer. *The McDonaldozation of Society*. Pine Forge Press, 1996. (Japanese Edition: K. Masaoka (1999). Waseda University Press).

T. Shibata. Mental commit robot for healing human mind. *Journal of the Robotics Society of Japan*, 17 (7):943–946, 1999. (in Japanese).

T. Tashima, S. Saito, T. Kudo, M. Osumi, and T. Shibata. Interactive pet robot with an emotional model. *Advanced Robotics*, 13(3):225–226, 1999.

S. Turkle. *Life on the Screen*. Simon & Schuster, 1995. (Japanese translation: M. Higure. Hayakawa–Shobo, 1998.).

S. Ukigaya. Bodies resisting medical discourses. *Gendai Shiso*, 28–10:132–152, 2000. (in Japanese).

TOWARD A METHOD FOR DETERMINING THE LEGAL STATUS OF A CONSCIOUS MACHINE

David J. Calverley J.D.

Research Fellow
Center for Law, Science and Technology
Arizona State University, Tempe, AZ USA
Davidc1016@aol.com

Abstract

As developers take the first tentative steps toward the creation of a machine consciousness, approaches embraced by terms such as "rational agents", "independent agents" or "autonomous agents" are viewed by some as necessary first steps. Central to these ideas is the notion that the "agent" will be "autonomous" in the sense that it is free of the direct constraints and explicit control of the developer. While the question of the legal liability of the designers of autonomous agents for the direct actions of such agents has been addressed to a limited extent in the relevant literature, there does not yet appear to have been an analysis of the criteria required to hold the machine consciousness itself legally responsible for its own actions. It is suggested that such an inquiry must examine the relationship between the philosophical ideas of consciousness and autonomy and legally relevant concepts such as personhood, and responsibility. This paper is an initial attempt to look at the concept of "responsibility" and determine if it may have any relevance to the concept of "autonomy". By examining the idea of responsibility in this way it is hoped that tentative steps will be outlined which will assist in the effort to integrate a machine consciousness, should it be developed, into the well developed fabric of modern jurisprudence. With such an end in mind, I discuss the risk that legally relevant factors may impinge upon the process, and could even prevent its coming to fruition. Finally I address the fact that the debate must, at some point in the future, shift from determining how and why a developer should be held accountable for the actions of the entity created, to an analysis of the factors which legally determine that the entity is truly independent, responsible for its own actions, subject to the same legal constraints affecting other members of the community, and able to lay claim to similar "rights" now available only to humans.

1. Introduction

> "Law is a socially constructed, intensely practical evaluative system of rules and institutions that guides and governs human action, that help us live together. It tells citizens what they may, must, and may not do, and what they are entitled to, and it includes institutions to ensure that law is made and enforced." (Morse 2004a page 158).

This definition, on its face, seems to be elegant and concise but, like an iceberg, it is deceptive. The tip only hints at the complexity of what is under the surface. Rather than simply applying the definition, let us begin by setting a foundation for the discussion which follows. In order to determine whether "law" has any normative value when it is used to evaluate the idea of machine consciousness we first need to gain at least a basic understanding of how this thing we call "law" is formulated at a conceptual level. By understanding what we mean when we speak of law, where it derives its ability to regulate human conduct, we can perhaps begin to formulate criteria by which some aspects of law could also be used to test the idea that something we have created in a machine substrate is a new form of conscious being. If this can be done in a way which is meaningful to both those who will be faced with deciding how to regulate such an entity and to the designers who are actually making the effort to create such an artifact, then it is worth the effort. As with most endeavors, it is often the question one asks at the outset which determines the nature of the debate and directs the form of the ultimate outcome. If we want to design a machine consciousness which we will

claim is the equivalent of a human, we should determine as early as possible in the process whether the result we seek will stand up to scrutiny and will, in the end, be amenable to being evaluated by criteria which are consistent with the way humans govern themselves and view each other.

2. Law as a Normative System

Applying the tools of analytic jurisprudence, philosophers of law struggle with the ways in which law is distinguishable from other normative systems. Historically, this has been done by conceptual analysis or intuition pumping with little or no empirical analysis. Only recently have empirical facts started to be used to evaluate legal concepts (Jones 1997; Leiter 1997). While acknowledging that there are many variations and nuances in legal theory, it is generally acknowledged that there have been two major historic themes which have, for the last few hundred years, dominated the debate about what "law" means.

One of the most familiar ideas to western societies is the concept of natural law, which was originally based on the Judeo-Christian belief that God is the source of all law. It was this belief which underpinned most of western civilization until the Enlightenment period. Prominent thinkers such as Augustine and Thomas Aquinas are two examples of this predominant orthodoxy. In essence, natural law proponents argue that law is inextricably linked with morality and therefore an 'unjust law is no law'.

With the Enlightenment came a decreasing emphasis on God as the giver of all law, and an increasing development of the idea that humans possessed innate qualities which gave rise to "law". As members of society, humans were capable of effecting their own decisions and consequently were entitled to govern their own actions based upon their intrinsic worth as individuals. While this concept was originally suggested by Hugo Grotius (1625) and later refined by John Locke (1739) it arguably reached its most notable actual expression in the system of laws ultimately idealized by the drafters of the United States Declaration of Independence. Drawing on a similar argument and applying it to moral philosophy, Emmanuel Kant hypothesized that humans were, by the exercise of their reason, capable of determining rules that were universally acceptable and applicable, and in turn able to use those rules to govern their conduct (Kant 1785).

More recently, John Finnis, building on ideas reminiscent of Kant, has outlined what he calls basic goods (which exist without any hierarchical ranking), and then has posited the existence of principles which are used to guide a person's choice when there are alternative goods to choose from. These principles, which he describes as the "basic requirements of practical reasonableness", are the connection between the basic good and ultimate moral choice. Derived from this view, law is the way in which groups of people are coordinated in order to effect a social good or to ease the way to reach other basic goods. Because law has the effect of promoting moral obligations it necessarily has binding effect (Finnis 1980). Similarly, Lon Fuller argued that law is a normative system for guiding people, and must therefore have an internal moral value in order to give it its validity. Only in this way can law fulfill its function which is to subject human conduct to the governance of rules (Fuller 1958; 1969). Another important modern theorist in this natural law tradition is Ronald Dworkin. Dworkin advocates a thesis which states in essence that legal principles are moral propositions grounded on past official acts such as statutes or precedent. As such, normative moral evaluation is required in order to understand law and how it should be applied (Dworkin 1978).

In contrast to the basic premise of natural law, that law and morality are inextricably intertwined, stands the doctrine of legal positivism. Initially articulated by Jeremy Bentham, and derived from his view that the belief in natural rights was "nonsense on stilts" (Bentham 1824), criticism of natural law centered around the proposition that law is the command of the sovereign, while morality tells us what law ought to be. This idea of law as a system of rules "laid down for the guidance of an intelligent being by an intelligent being having power over him", was given full voice by Bentham's protégé, John Austin. In its simplest form this idea is premised on the belief that law is a creature of society and is a normative system based upon the will of those ruled as expressed by the sovereign. Law derives its normative power from the citizen's ability to know and

predict what the sovereign will do if the law is transgressed (Austin 1832).

Austin's position, that law was based on the coercive power of the sovereign, has been severely criticized by the modern positivist H.L.A. Hart who has argued that law requires more than mere sanctions; there must be reasons and justifications why those sanctions properly should apply. While neither of these positions rule out the overlap between law and morality, both do argue that what constitutes law in a society is based on social convention. Hart goes further and states that this convention forms a rule of recognition, under which the law is accepted by the interpreters of the law, i.e. judges (Hart 1958; 1961). In contrast, Joseph Raz argues that law is normative and derives its authority from the fact that it is a social institution which can claim legitimate authority to set normative standards. Law serves an essential function as a mediator between its subjects and points them to the right reason in any given circumstance, without the need to refer to external normative systems such as morality (Raz 1975).

It is conceded that the above exposition is vastly over simplified and does not do justice to the nuances of any of the described theories. None the less, it can serve as a basis upon which to premise the contention that despite the seeming difference between the two views of law, there is an important point of commonality. Returning to the definition with which we started this paper, we can see that it is inherently legal positivist in its outlook. However, its central idea, that law is a normative system by which humans govern their conduct, seems to be a characteristic shared by both major theories of law and therefore is one upon which we can profitably ground some further speculation. To the extent that law requires humans to act in accordance either with a moral norm established in accordance with a theological, or natural theory, or to the extent it is a normative system based on one's recognition of and compliance with a social created standard of conduct, it is premised on the belief that humans are capable of, and regularly engage in, independent reflective thought, and are able to make determinations which direct their actions based upon those thoughts. Described in a slightly different way, law is based on the premise that humans are capable of making determinations about their actions based on reason.

> "Human action is distinguished from all other phenomena because only action is explained by reasons resulting from desires and beliefs, rather than simply by mechanistic causes. Only human beings are fully intentional creatures. To ask why a person acted a certain way is to ask for reasons for action, not the reductionist biophysical, psychological, or sociological explanations. To comprehend fully why an agent has particular desires, beliefs, and reasons requires biophysical, psychological, and sociological explanations, but ultimately, human action is not simply the mechanistic outcome of mechanistic variables. Only persons can deliberate about what action to perform and can determine their conduct by practical reason" (Morse 2004a page 160).

Similarly, Gazzaniga and Steven (2004 page 67), express the idea as follows:

> "At the crux of the problem is the legal systems view of human behavior. It assumes (X) is a "practical reasoner", a person who acts because he has freely chosen to act. This simple but powerful assumption drives the entire legal system.
> …
> The view of human behavior offered by neuroscience is simply at odds with this idea. …neuroscience is in the business of determining the mechanistic actions of the nervous system. The brain is an evolved system, a decision-making device that interacts with its environment in a way that allows it to learn rules to govern how it responds. It is a rule based device that, fortunately, works automatically.
> …
> Neuroscience will never find the brain correlate of responsibility, because that is something we ascribe to humans, not to brains. It is a moral value we demand of our fellow, rule-following human beings. … The issue of responsibility … is a matter of

> social choice. …We are all part of a deterministic system that some day, in theory, we will completely understand. Yet the idea of responsibility is a social construct and exists in the rules of society. It does not exist in the neuronal structure of the brain."

The point is clear that the law looks at the motivation of the actor and the ability of the actor to control actions based upon those motivations. Gazzaniga has expressed similar arguments as follows:

> "Neuroscience seeks an empirically valid model of human nature and human behavior – one that has predictive power and allows us to understand better the relation between our brains and mental lives. The law seeks to bring about conformity of individuals' behavior to certain codes in order to maintain order in society" (Waldbauer 2001 page 364).

While this perspective is not universally accepted by philosophers of law (Goodenough 2001), for the purpose of this paper it has been used as the basis from which to argue that proponents of machine consciousness have significant hurdles to overcome to prove an assertion that a machine consciousness can be seen to be a legally responsible entity. Interestingly enough, it is possible that while this view of law may affect machine consciousness design, it is equally likely that if the design is ultimately successful, we may have to revisit some of the basic premises of law.

Why take this exposition as the starting place of our analysis? The answer is rather straight forward. Gazzaniga's and Morse's presentations were commissioned for a conference organized by the American Association for the Advancement of Science on Neuroscience and Law. In September 2004, Morse presented his ideas to the President's Council on Bioethics. Gazzaniga is a member of the President's Council. Consequently, it seems to me that the ideas expressed are positioned in such a way that they can uniquely influence the future course of deliberations at the highest level of government, at least in the United States. By stating that neuroscience can never affect law because law is based on a humanly constructed concept of responsibility, a bias is created against any reductive material theory which argues to the contrary. As stated earlier, how the question is framed often defines the answer received. My purpose here is to change the frame of reference slightly and rather than look at how things exist in the world today, thereby forcing the debate into the terms set forth by Morse, change the reference to ask whether we can, in the present state of knowledge, posit a scenario where the skepticism expressed can be tested not by looking backward in evolutionary time and trying to decide why humans have laws and chimpanzees don't (Morse 2004b), but rather to look forward and to determine if there is a set of conditions which, if they came to pass, would plausibly require us to reevaluate our position in the world and the relevance of law to that position.

3. Law and Machine Consciousness

Let us look at how our view of law as a normative system based upon responsibility relates to machine consciousness. The debate concerning the mind body problem and its nuances is beyond the scope of this paper. Likewise, "consciousness", a topic of much recent scholarship, can only be dealt with in a cursory manner. Suffice it to say that the comments I am making are based on a superficial sketch of the issues and by no means fully describe the various controversies surrounding these concepts. However, I believe that further analysis of the points outlined in the context of legal systems may lead to useful insights. Certainly the central question here is what is meant by consciousness, but again, while we need to define the term in a meaningful way we do not need to solve the problem of what it will ultimately come to mean.

A threshold question is whether we are talking about functional consciousness or phenomenal consciousness. In the first instance, the fact that an artifact looks like a duck, talks like a duck, and walks like a duck, leads to the conclusion that it must be a conscious duck. In the second instance, phenomenal consciousness, there must be more: the artifact must not only think it is a duck, it must feel internally that it is "duckie". Looking at the various indicia of consciousness (Torrance 2004), it is possible to argue that there are degrees of consciousness

some of which are less than would be required for ascription as "full" human beings. Certainly if one looks at the question of moral rights, setting aside the question of law for the moment, we see many instances, such as a fetus or people in persistent vegetative states, where biological beings exhibiting less than all indicia of consciousness, are none the less ascribed certain rights.

As a brief aside, an area where additional analysis is required relates to analogous intermediate stages in the development of a machine consciousness. As we move along the path toward a machine consciousness, legitimate arguments can be made that the endeavor itself is replete with moral and ethical pitfalls. If the same logic as urged for animal rights, or for the rights of fetuses, is applied to a machine consciousness, some of these issues could have the potential to radically curtail actual development of a conscious entity. If part of the process of developing a machine consciousness is an emergent learning type process, or even a process of creating various modules which add attributes of consciousness such as sentience, nociception (sensing pain stimuli), or language, in a cumulative fashion, some could argue that this is immoral. As posed by LeChat (1986 page 75, 78), the question becomes: "Is the AI experiment then immoral from its inception, assuming, that is, that the end (telos) of the experiment is the production of a person? An AI experiment that aims at producing a self-reflexively conscious and communicative "person" is *prima facie* immoral." In the present day we are all aware of the constraints on both human experimentation and the use of higher primates in various types of medical procedures. Must designers of a machine consciousness be aware that as they come closer to their goal, they may have to consider such protocol in their experimentation? Arguably yes, if human equivalence is the ultimate goal. Failure to treat a machine consciousness in this way could be viewed as a form of speciesism. The utilitarian philosopher JJC Smart (1973) has observed "…if it became possible to control our evolution in such a way as to develop a superior species, then the difference between species morality and a morality of all sentient beings would become much more of a live issue." Interestingly, this was written before Peter Singer (1975), the premier animal rights advocate, presented the term "speciesism". It is worthwhile noting in this context, that following a strict interpretation of classical natural rights theory would lead to a conclusion that the creation and exploitation of a machine consciousness is allowed, so long as the machine consciousness is viewed as property and safeguards can be installed to assure that it does not become a "runaway." On the other hand, and perhaps more in accord with modern thinking, if we assume that the machine consciousness has sufficient capability, either inherently, or because it is part of a group where the average mature individual has those characteristics, then it is conceivable to view the rights more from the perspective of liberal individualism and look to ascribe the machine consciousness, as an individual, with "natural rights". The logical extension of this position is to ascribe rights simply by virtue of the fact of the existence of a machine consciousness. In turn this attribution brings us back to LeChat's assertion of *prima facie* immorality (Calverley 2003). I suggest that much more work, both at a theoretical level and a practical level, needs to be done in this area in order to safeguard against the possibility that any endeavor to develop a machine consciousness will be terminated on moral and ethical grounds before it can be empirically tested (Calverley 2004).

Let's return to the question of legal perspective. If one has a bias toward legal positivism, then a likely outcome, if faced with the actuality of a machine consciousness, is to simply say that society will decide, hopefully after full and open debate taking into account the moral issues just mentioned, whether to ascribe a particular status to such an artifact. This response, much like the positions taken in the animal rights debate, is for the most part utilitarian based and pragmatic in its orientation. On the other hand if, as just noted, one has a bias toward a more natural law view, it is possible to argue that a machine consciousness has intrinsic rights. In either case we are left with the question of whether there is some legally relevant characteristic which, as Morse and Gazzaniga assert, is presently viewed as purely human but which we can identify and articulate in such a way as to say that if that same characteristic is exhibited or possessed by another entity then that entity is entitled to be treated in an equivalent manner to humans.

4. Responsibility and Autonomy

If asked whether humans are different from animals most people would say yes. When pressed to describe what that implies in the context of legal rules, many people would respond that it means we have free will, that our actions are not predetermined. Note however that Morse (2004a) argues that this is a mistake in that free will is not necessarily a criteria for responsibility in a legal sense. From the perspective of moral philosophy the debate can be couched in slightly different terms. In the view of the "incompatibilist", in order for people to be held responsible for their acts they must have freedom to choose amongst various alternatives. Without alternatives there can be no free will (vanInwagen 1986; Kane 1996). The incompatibilist position has been strongly attacked by Harry Frankfurt, who called their argument the "principle of alternate possibilities". Frankfurt has argued that it is possible to reconcile free will with determinism in his view of "personhood". His conclusion is that people, as opposed to animals or other lower order beings, possess first and second order desires as well as first and second order volitions. If a person has a second order desire it means that she cares about her first order desires. To the extent that this second order desire is motivated by a second order volition, that is, wanting the second order desire to be effective in controlling the first order desire, the person is viewed as being autonomous so long as she is satisfied with the desire. The conclusion is that in such a case the person is autonomous.

It should be noted that in this context Frankfurt is using the term person as the equivalent of human. Others would argue that person is a broader term and more inclusive, drawing a clear distinction between person and human (Strawson 1959; Ayers 1963). My preference is to use the term human to apply to homo-sapiens and the term person to conscious beings irrespective of species boundaries. However, much theoretical analysis remains to be done on this distinction and the moral implications which arise from it, particularly in the context of machine consciousness.

It is helpful in this regard to compare Frankfurt's position with Kant's belief that autonomy is viewed as obedience to the rational dictates of the moral law (Hermann 2002). Kant's idea that autonomy is rational also differs from that of David Hume (1739) who argued that emotions are the driving force behind moral judgments. Hume seems to be an antecedent of Frankfurt's concept of "satisfaction" if the latter's essay on love is understood correctly (Frankfurt 1994). Transposing these contrasting positions into the language used earlier to describe law, I suggest that it is possible to equate this sense of autonomy with the concept of responsibility. Humans are believed to be freely capable of desiring to choose and actually choosing a course of action. Humans are believed to be capable of changing desires through the sheer force of mental effort applied in a self reflexive way. Humans are therefore, as practical reasoners, capable of being subject to law so long as they act in an autonomous way.

Autonomy however, has a number of potential other meanings in the context of machine consciousness. Consequently, we need to look at this more closely if we are to determine whether the above discussion has any validity in aiding the design of a machine consciousness.

Hexmoor et. al., (2003), draw a number of distinctions between the different types of interactions relevant to systems design and artificial intelligence. First there is human to agent interaction where the agent is expected to acquire and conform to the preferences set by the human operator. In their words, "(a) device is autonomous when the device faithfully carries the human's preferences and performs actions accordingly." Another sense is where the reference point is another agent rather than a human. In this sense the agents are considered relative to each other and essentially negotiate to accomplish tasks. In this view "(t)he agent is supposed to use its knowledge, its intelligence, and its ability, and to exert a degree of discretion." In a third sense there is the idea mentioned before that the agent can be viewed as manipulating "…its own internal capabilities, its own liberties and what it allows itself to experience about the outside world as a whole." Margaret Boden, in a similar vein, writes about the capacity of the agent to be original, unguided by outside sources (Boden 1996). It is in this third sense where I suggest that the term autonomy comes closest to what the law views as crucial to its sense of responsibility. However, before exploring that point further, I believe that a digression to show how the first

two of the alternative senses of autonomy mentioned can readily be accepted and dealt with by the legal system.

In each of the first two senses of autonomy discussed above, there appears to be a referent to which the agent always defers in making its decision. In the first sense it is the human operator. Similarly in the second sense, it is the weighting or value placed upon the various decisions, weighting which is determined, not by the agent but by the operator who is setting the conditions for the agent's interactions with other agents. In each of these situations there appears to be a "controlling" entity which is setting the parameters of action. From a philosophical and legal sense this would strongly imply that the agent is not the competent causal agent of a consequence that has legal significance.

Let's look at this more closely. Assume for example that I program an agent so that it enters virtual space, say the Internet, to perform a task I set for it, say, locating a particular set of documents. I give it various search criteria, i.e., set the search parameters, then leave it to its own devices in determining how best to accomplish the tasks and fulfill its duty. Assume further that the agent in fact proceeds to do as directed, but in the process commits what you and I and the world would view as an egregious harm. Perhaps in order to get the document it has to fraudulently represent to another agent that it is authorized to access a particular computer. Perhaps it determines that the best way to obtain the document requested is to copy if from a site where there is a charge for access. In order to avoid this fee, it manipulates another computer to access a third person's bank account to make the payment. Because the initial directions did not explicitly rule out these courses of action, the artifact is not constrained from following them.

In each of these cases the law would have little difficulty in ignoring the "autonomy" of the agent and ascribing legal responsibility to the person who programmed the computer. As explained in Heckman (1999), the law, using various well established rules such as strict products liability, and others, would have little difficulty in determining that the real actor in this scenario is the person who sets the chain of action into motion.

As a further aside, space considerations preclude analysis of the meaning of the term "agent" as it is used in philosophy and in law. Suffice it to say that law has a technical understanding of the term agent which implies that the agent is directed and controlled by a principal which may, if one is not careful, predispose one to conclude that the first two senses of autonomy are the only legally relevant ones (Restatement, Agency). Another topic beyond the scope of this paper, which provides a basis for speculation, is whether a philosophical agent can act morally without exhibiting free will, mental states or responsibility (Floridi 2004).

In the third sense of autonomy mentioned above, the answer is not so straight forward. Change the above scenario slightly and assume that our initial point of departure is merely a stated desire we have to read a particular document. Our "friend", a conscious machine, hears our expression, and motivated by friendliness and social convention, decides to get the document for us as a birthday present. Acting upon this determination the machine consciousness then proceeds to commit similar proscribed acts as mentioned above. Here, I suggest, something more has happened, something more human like. If, in this scenario, the agent is autonomous in the sense described by Frankfurt, what I call the strong sense of autonomy, then it is conceivable to say that the law could directly affect the question of how we effectively evaluate a machine consciousness. If we adopt the strong definition of autonomy, and argue that if it is achieved in a machine, as it would be in the above example, then at least from a functional viewpoint, we could assert the machine is the equivalent of a human in terms of its being held responsible. But one could easily be faced with the objection that such a conclusion simply begs the question about whether the artifact is phenomenally conscious (Adams 2004). It is in addressing this point that our concept of law as a normative system can perhaps guide us to a deeper understanding of autonomy.

In its simplest form, to be phenomenally conscious means to know what it is like to be something. Taken in this sense, we can see that this looks a lot like Frankfurt's person who is capable of forming second order volition. In our example, if one can conceive of a second order volition, the desire to be a good friend and to

comply with social convention, and can as a result affect a first order action, the obtaining of the document, constrained only by the idea that one is satisfied by that result, does that not imply phenomenal consciousness? Going the next step then, we can argue that law acts at the level of this second order volition. It sets parameters which, as society has determined, outline the limits of an accepted range of responses within the circumscribed field which it addresses, say contract law or tort law or criminal law. This would imply that law acts in an exclusionary fashion in that it inhibits particular first order desires and takes them out of the range of acceptable alternatives for action (Raz 1975: 1986; Green 1988). Note that this does not mean to imply that these are the only possible responses or even the best responses the actor could make. To the extent that the subject to which the law is directed, (the citizen within the control of the sovereign in Austin's terms) has access to law as normative information, she can order her desires or actions in accordance with law or not. This would mean, to borrow the terminology of Antonio Damasio (1994), that the law sets the somatic markers by which future actions will be governed. I suggest that this does not require that the artifact have a universal, comprehensive understanding of the law any more than the average human does. Heuristics, or perhaps concepts of bounded rationality, could provide the basis for making decisions which are "good enough" (Clark 2003). Similar arguments have been advanced on the role of emotion in the development of a machine consciousness (Sloman 1981; Arbib 2004; Wallach 2004). Perhaps, in light of work being done in how humans make decisions (Kahneman 1982; Lakoff 1987), more pointed analysis is required to fully articulate the claim concerning law's normative role within the context of autonomous behavior. One further caution, even though I suggest that accepting law as a guide to a second order volition does not diminish the actor's autonomy, this proposition can be challenged by some theories such as anarchism (Wolff 1970).

Ultimately the question comes down to whether this type of second order volition can be instantiated in a machine consciousness or whether it is exclusively the realm of biological species. John Searle (1980) and others would say that it can only be accomplished in biological systems. Others however, have started to look at theoretical possibilities where just this type of activity can occur in non-biological systems (Covigaru 1991; Clark 2003; Holland 2003). In actively moving toward a machine consciousness, it is possible that this question will be answered empirically.

5. Conclusions

What then can we make of this argument? I believe that if a claim of autonomy in Frankfurt's sense could plausibly be made for a machine consciousness, and could therefore show that this characteristic is no longer uniquely human, it is equally plausible to argue that responsibility for action can shift from the developer to the machine consciousness, thereby making the artifact a moral agent not simply a moral patient. Next, I suggest that if we can plausibly present empirical evidence that a machine consciousness can in fact develop second order volitions basing them upon the normative characteristics of law, then it is conceivable and logical to argue that it is in fact capable of conforming its "will" to law, independent of the form of its initial state of programming. In this regard we still have to struggle with the first party-third party problem of consciousness, but perhaps the fact that the process of development of a machine consciousness is more transparent than in biological system will assist in this regard. Finally I argue that while this will not mean that law is automatically applicable to such an entity, it will mean that at some point in time, the law will have to accommodate such an entity (Solum 1992), and in ways which could force humans to re-evaluate their concepts of themselves. If such a machine consciousness existed, it would be conceivable that it could legitimately assert a claim to a certain level of rights which could only be denied by an illogical assertion of species specific response.

Looking at law as a humanly devised normative concept, and in particular focusing on the idea of responsibility, we can assert that nothing we have seen will necessarily preclude a machine consciousness from acceptance as a legally competent actor. On the contrary, the idea of responsibility sharpens and focuses the question of what it means to be human in a unique way which designers of a machine consciousness can effectively apply.

References:

Adams, W. 2004: "Machine Consciousness: Plausible Idea or Semantic Distortion?" 11 J. of Consciousness Studies, No. 9, Sept. 2004

Arbib, M. and Fellous, J. 2004: "Emotions: from brain to robot" 8 Trends in Cognitive Science No. 12 page 554

Austin, J. 1832: *The Province of Jurisprudence Determined* (1955 edn.) London: Weidenfeld and Nicholson.

Ayers, A.J. 1963: *The Concept of a Person* New York: St. Martin's Press

Bentham, J. 1824: "Anarchical Fallacies" in *The Works of Jeremy Bentham* vol. 2, J. Bowring, (1962) New York: Russell and Russell

Boden, M. 1996: "Autonomy and Artificiality" in *The Philosophy of Artificial Life* (ed. Boden, M.) Oxford: Oxford Univ. Press

Calverley, D. 2003: "Imagining Rights for an Artificial Intelligence", unpublished paper presented at "Transvision 2003", Yale University, June 2003.

Calverley, D. 2004: "Ethical Implications of Artificial Intelligence Design" in *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and Artificial Intelligence Vol. III,* (ed. Smit, I., Wallach, W. and Lasker, G.) Windsor Canada: IIAS

Clark, A. 2003: "Artificial Intelligence and the Many Faces of Reason," in Stich, S. and Warfield, T. (eds.) *The Blackwell Guide to Philosophy of Mind* Malden, MA: Blackwell Publishing

Covigaru, A. and Lindsay, R. 1991: Deterministic Autonomous Systems," AI Magazine, Fall 1991 page 110.

Damasio, A. 1994: *Descartes Error* New York: Quill paperback, Harper Collins Publisher

DeGrazia, D. 1996: *Taking Animals Seriously* New York: Cambridge Univ. Press

Dworkin, R. 1978: *Taking Rights Seriously* (revised edn.) London: Duckworth

Finnis, J. 1980: *Natural Law and Natural Rights.* Oxford: Clarendon Press.

Floridi, L. and Sanders, J.W. 2004: "On the Morality of artificial agents," 14 Minds and Machines vol. 3 at 349: available online at http://www.wolfson.ox.ac.uk/~floridi/

Frankfurt, H. 1969: "Alternate possibilities and moral responsibility," in *The Importance of What We Care About* Cambridge: Cambridge Univ. Press (1988).

Frankfurt, H. 1971: "Freedom of the will and the concept of a person," in *The Importance of What We Care About* Cambridge: Cambridge Univ. Press (1988)

Frankfurt, H. 1994: "Autonomy, Necessity and Love," in *Necessity, Volition and Love* Cambridge: Cambridge Univ. Press (1999).

Fuller, L. 1958: "Positivism and Fidelity to Law - a response to Professor Hart," 71 Harvard L. Rev. 630 (1958)

Fuller, L. 1969: *The Morality of Law* (2nd. edn.) New Have: Yale University Press

Gazzaniga, M. and Steven M. 2004: "Free Will in the Twenty-First Century: A Discussion of Neuroscience and the Law," in *Neuroscience and the Law* New York: Dana Press

Goodenough, O. 2001: "Mapping Cortical Areas Associated with Legal Reasoning and Moral Intuition," 41 Jurimetrics J. 429 (2001)

Green, L. 1988: *The Authority of the State* Oxford: Clarendon Press

Grotius, H. 1625: *De Jure Belli ac Pacis Libri Tres*, (tr. F. Kelsen) Oxford: Clarendon Press

Hart, H.L.A. 1958: "Positivism and the separation of law and morals," 71 Harvard L. Rev. 593 (1958)

Hart, H.L.A. 1961: *The Concept of Law* Oxford: Clarendon Press

Heckman, C. 1999: "Liability for Autonomous Agent Design," in *Autonomous Agents and Multi-agent Systems* The Netherlands: Kluwer Academic Publishers

Herman, B. 2002: "Bootstrapping" in *Contours of Agency* (ed. Buss, S. and Overton, L.) Cambridge, Mass: The MIT Press

Hexmoor, H., Castelfranchi, C., and Falcone, R. 2003: "A Prospectus on Agent Autonomy," in *Agent Autonomy* (ed. Hexmoor, H. et. Al.) Boston: Kluwer Academic Publishers

Holland, O (ed.) 2003: "Machine Consciousness" 10 J. of Consciousness Studies, No 4-5, April/May 2003.

Hume, D. 1739: *A Treatise of Human Nature* (ed. P. Nidditch 1978) Oxford: Clarendon Press

Jones, O. 1997: "Evolutionary Analysis in Law: An Introduction and Application to Child Abuse," 75 N.C. L. Rev. 1117 (1997).

Kahneman, D., Slovic, P. and Tversky, A. (eds.) 1982: *Judgment under Uncertainty: Heuristics and Biases* Cambridge: Cambridge Univ. Press

Kane, R. 1996: *The Significance of Free Will* New York: Oxford Univ. Press

Kant, E. 1785: *Grounding of the Metaphysics of Morals* (tr. J. Ellington 1981) Indianapolis: Hackett

Lakoff, G. 1987: *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind* Chicago: University of Chicago Press

LeChat, M. 1986: "Artificial Intelligence and Ethics: An Exercise in Moral Imagination" AI Magazine, Summer 1986 page 70

Leiter, B. 1997: "Rethinking Legal Realism: Toward a Naturalized Jurisprudence," 76 Texas L. Rev. 267 (1997).

Locke, J. 1739: "Two Treatises of Government" in *Two Treatises of Government: a critical edition* (1967) London: Cambridge Univ. Press

Morse, S. 2004a: "New Neuroscience, Old Problems," in *Neuroscience and the Law* New York: Dana Press

Morse, S. 2004b: Presentation to President's Council on Bioethics, September 9, 2004, available online at http://bioethicsprint.bioethics.gov/transcripts/sep04/session1.html.

Raz, J. 1975: *Practical Reason and Norms* London: Hutchinson

Raz, J. 1986: *The Morality of Freedom* Oxford: Clarendon Press

Restatement of the Law, Second, Agency: Philadelphia: American Law Institute

Searle, J. 1980: "Minds, Brains and Programs," 3 Behavioral and Brain Sciences 417

Singer, P. 1975: *Animal Liberation* New York: Random House

Sloman, A and Croucher, M. 1981: "Why Robots will have Emotions," in *Proceedings IJCAI 1981*, Vancouver

Solum, L. 1992: "Legal Personhood for Artificial Intelligences" 70 North Carolina L. Rev. 1231

Strawson, P. 1959: *Individuals* London: Methune

Torrance, S. 2004: "Us and Them: Living with self-aware systems," in *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and Artificial Intelligence Vol. III,* (ed. Smit, I., Wallach, W. and Lasker, G.) Windsor Canada: IIAS

Van Inwagen, P. 1983: *An Essay on Free Will* Oxford: Oxford Univ. Press

Waldbauer, J. and Gazzaniga, M. (2001): "The Divergence of Neuroscience and Law" 41 Jurimetrics J. 357 (2001)

Wallace, W. 2004: "Artificial Morality: Bounded Rationality, Bounded Morality and Emotions," *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and Artificial Intelligence Vol. III,* (ed. Smit, I., Wallach, W. and Lasker, G.) Windsor Canada: IIAS

Wolff, R.P. 1970: *In Defense of Anarchism* (revised 1998) Berkeley and Los Angeles: University of California Press

An Ordinal Probability Scale for Synthetic Phenomenology

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

Abstract

Synthetic phenomenology can be broken down into three areas: (1) the determination whether a system is capable of phenomenal states, (2) the identification of the mental content of the machine (the machine's conceptual and non-conceptual representations), and (3) the analysis of a particular structure of mental content to identify the parts that are phenomenally conscious. This paper proposes that an ordinal probability scale could be used to address the first of these problems and sets out a proposal for such a scale that ranks machines according to the likelihood that they are capable of experiencing phenomenal states. The overall approach suggested here will be used to describe the synthetic phenomenology of Holland's and Troscianko's 'conscious' robot that is currently under development at the University of Essex and the University of Bristol.

1 Introduction

Research on machine consciousness aims to develop machines that exhibit conscious behaviour and might be capable of phenomenal states. A description of machines' phenomenal states is provided by synthetic phenomenology, which attempts to discover whether robots can have conscious experiences and to articulate them when they occur.

One of the challenges of synthetic phenomenology is that it seems possible that a zombie robot could perfectly mimic human behaviour without experiencing anything at all, and so external behaviour does not seem to be a reliable guide to conscious states. It might be thought that we could solve this by looking at the internal structure of the robot. If the robot contains structures that are correlated with or cause consciousness in humans, then it is likely to experience phenomenal states. The first half of this paper covers the problem with this approach: With only behavioural evidence to go on it is impossible to empirically identify a necessary and sufficient set of the correlates or causes of consciousness. Without this, we cannot identify what needs to be included in a machine to make it conscious and cannot tell whether a machine that exhibits conscious behaviour is likely to be experiencing phenomenal states.

To address this difficulty, the second half of this paper sets out a proposal for an ordinal probability scale that ranks machines according to the likelihood that they are capable of sustaining phenomenal consciousness, based on their proximity to our own case. The factors that are used in this scale are the rate of information processing, the size of the machine, the way in which its sub-functions are assembled to produce the global functions (functional granularity), the machine's time-slicing and whether it is analogue or digital. Weightings are given to each of these factors and the combination of these weights is used to situate each machine on an ordinal scale.

This scale is put forward as a pragmatic tool that will enable us to proceed with synthetic phenomenology and address some of the ethical issues in machine consciousness. At this stage, the proposed factors, along with their weightings, should be seen as tentative first suggestions, which I hope will be criticised and develop in the longer term into a commonly agreed scale.

2 Synthetic Phenomenology

Whilst synthetic phenomenology can be used to refer to the *synthesizing* of phenomenal states (Jordan (1998) coined the term in this connection), it can also be used to describe the *phenomenology* of artificial systems that may or may not be experiencing conscious states. It is in the latter sense that I will be using it in this paper. Husserl's phenomenological project was the description of human consciousness (without any commitments to the natural attitude); the synthetic phenomenological project is the description of machine consciousness. Synthetic phenomenology is a way in which people working on machine consciousness can measure the

extent to which they have succeeded in realising consciousness in a machine.

The phenomenology of artificial systems can be broken down into three stages: (1) the determination whether a system is capable of phenomenal states, (2) the identification of the 'mental content' of the machine[1] (the machine's conceptual and non-conceptual representations), and (3) the analysis of a particular structure of mental content to identify the phenomenally conscious parts. The first of these stages is the main focus of this paper, but before examining it in detail I will give a brief overview of all three stages in order to clarify the relationship between them.

2.1 Can a machine experience phenomenal states?

The external behaviour of a robot could be taken to indicate that it is experiencing phenomenal states, or it could just be the result of unconscious automatic processing. Even if the robot could master language and pass the Turing test, this would not guarantee that it is experiencing phenomenal qualia. It is not even inconsistent or wrong to see animals or other human beings as automatons.

We avoid solipsism and attribute consciousness to other humans and some animals because we share a similar biology. In the case of machines, this common underlying substrate is missing and so we need to find some other way to decide whether they are capable of phenomenal states or not. The most promising line of approach would be to work out what it is about our biology that makes us conscious – its proteins, neurons, functions or representations, for example - and then look inside the robot to see if these consciousness-producing properties are present as well. Alternatively, we could look for the *correlates* of phenomenal states in the brain. This weaker approach looks for the factors that are present whenever consciousness is present without trying to explain how these factors actually lead to conscious states. When we have identified either the mechanisms in the brain that produces phenomenal states or the correlates of phenomenal states, we can see if these mechanisms or correlates are present in the machine. If they are, then it seems reasonable to conclude that the machine may be capable of phenomenal states.

Unfortunately it does not look as if either of these approaches will be able to answer the question about the consciousness of non-biological entities. At present we have no idea about how the brain produces conscious states and there are some potentially irresolvable difficulties with empirically separating out the correlates of consciousness, which may prevent us from making any progress at all in this direction. Section 3 covers these problems in detail. If a solution cannot be found, we may only be able to answer the questions raised by this first proposed stage of synthetic phenomenology with a probabilistic assessment of the likelihood that a machine can support phenomenal states. This will be discussed in the second half of this paper.

2.2 Identification of the mental content of a machine

As a machine interacts with its real or virtual environment it processes sense data into some kind of representation of its world – that there is a green apple five metres from its hand, that its hand is clenched, that it is feeling sad, and so on. These internal representations are the mental content of the machine.

The identification of conscious mental content in humans is imperfect, but relatively straightforward since we share language and a common biological base. When people describe their phenomenal world our common biology leads us to assume that it is very similar to our own. However, this is not the whole story, because people also have a great deal of unconscious and non-conceptual mental content that can only be identified indirectly (Chrisley, 1995).

The identification of mental content in robots is more challenging because they generally have rudimentary language and are built in a very different way from humans. This means that we cannot simplistically assume that their phenomenal experiences are in any way similar to our own. When the robot's information-processing is based around neural networks, we can try to identify the robot's mental content by exposing it to stimuli and interpreting the active parts of the network as representations of the external stimulus. By systematically varying the stimuli, a map of the representations in the network can be worked out and used to describe the robot's mental content in novel situations.

A second technique was suggested by Holland and Goodman (2003). In their experiments a simple robot was programmed to move around its environment and build up 'concepts' corresponding to combinations of sensory input and motor output. Once the robot's concept formation was complete, its mental content could be identified by a process of 'inversion'. Each concept was a combination of data about the distance of environmental features and information about how the robot moved during the sample period. The inversion consisted in plotting the movements and distances recorded by each

[1] I will be using the term 'mental content' to refer to a machine's representations and the non-conceptual content of its 'mind'.

concept to generate a map of the robot's representation of its environment.

2.3 Separation of conscious from unconscious mental content

Machines that can support phenomenal states present a third problem for synthetic phenomenology. If they are anything like humans, it is likely that at any point in time some of their mental content is conscious and the rest unconscious. In the well-known example of driving home from work, a person can avoid obstacles in the road, stop at traffic lights, navigate perfectly and yet see and remember nothing of the journey because they are preoccupied with something else. The sense data from the road, steering wheel and pedals must be mental content in some form, but it is not conscious mental content. Some feature of the mental content that the person is attending to must differentiate it from information about the road, steering wheel and pedals.

Although robots might be *capable* of consciousness, at any point in time it is possible that none of their mental content is actually conscious, and one of the tasks of synthetic phenomenology is to distinguish the conscious from the unconscious mental content. Different theories about human consciousness are gradually converging around the idea that it is the structure and interconnection of information that makes the difference between conscious and unconscious content (See Dehaene (2001) for an overview). To carry out this part of its task, synthetic phenomenology can use these theories about consciousness (for example, the global workspace model put forward by Baars (1988) or Metzinger's (2003) constraints) to identify the phenomenal mental content of the machine.

3 Inferring Consciousness from Behaviour

3.1 Local and global functions

The brain carries out a wide variety of functions, ranging from information and language-processing to low level functions in the neurons' ion channels. Some of these functions are *local* to regions of the brain, for example transforming retinal data entering V5 into movement information, whereas other functions are *global* to the whole brain, for example transforming incoming sensory data about an apple into an output instructing the arm to pick it up.

It might be thought that we can easily determine which of the brain's local functions are necessary and sufficient for consciousness. For example, if we damage the function of V5, then the person loses consciousness of movement information (see Zihl et al. (1983) for a case study and also Zeki and Bartels (1998) for the notion that micro-consciousnesses are distributed throughout the brain). However, this type of experiment only identifies a link between a local function and consciousness *indirectly* through its impact on the brain's global functions. If the person's global functions were not affected by damage to V5, we would have no idea whether the local function carried out by it had any effect on consciousness.

It is even harder to decide whether the way in which a global or local function is implemented affects consciousness. The brain's global or local functions could be carried out by neurons or the population of China, but as long as its global functions remain constant, it will always describe its conscious experiences in the same way. A function that was carried out consciously when there were biological neurons present might be carried out unconsciously when there are no biological neurons present, but the function is still carried out, and so in both cases the person will continue to respond to the input: "Are you conscious?" with the output "Yes!" even if there is no longer any consciousness present. To make this point clearer I will consider a thought experiment that is often discussed in the literature, in which part of a person's brain is replaced by a chip that carries out the same functions as the brain part that is replaced.

3.2 Silicon brain functions

At first glance the replacement of a brain part by a chip seems to hold out the prospect of identifying whether the way in which the brain's functions are implemented affects consciousness. If we replaced V5 with a functionally equivalent chip and lost consciousness of movement information, then we could conclude that the brain's biological substrate and functions are *both* necessary for consciousness. However, as Moor (1988) and Prinz (2003) point out, since the global behaviour of the person would not be changed by the operation, neither an external observer nor the person who received the chip implant could observe any effect of the replacement on consciousness.

An outside observer would not detect the replaced part because the function of V5 would still be carried out by the chip. The person would still report movement information that is processed by affected area, even though there may not be any consciousness of movement present. From an outside point of view, this will not even seem like a confabulation because the visual system will be working perfectly.

A first-person perspective does not help matters either. Since the chip is functionally connected to

the rest of the brain in the same way that V5 was before the operation, our language centres will report phenomenal movement in the same way that they did before and it has already been established that the external behaviour of the person will remain unchanged. Searle (1992, 66-7) thinks that we might feel forced to say that we experience movement even though we do not experience any movement. However, if this distinction between inner thought and outer behaviour was conscious and could be remembered, it could be reported at a later time, and so there would be a change in the subject's behaviour, which is ruled out by this experiment. Furthermore, as Moor points out, the chip must also give the person the *belief* that they are conscious of the functions processed by the chip and so Searle cannot experience one thing and believe another. It seems that even a first-person perspective cannot be used to decide whether consciousness is affected by the replacement of biological neurons with a functionally equivalent chip.

Against this Chalmers (1996) argues that verbal behaviour and consciousness would be very tenuously connected if we could lose our conscious experience of movement and yet continue to describe movement using language. The problem with this objection is that the implantation of a chip involves invasive surgery and it is not uncommon for people with brain damage to be systematically mistaken about their experiences and confabulate to an extraordinary extent to cover up their deficiency. For example, people with Anton's syndrome are blind and yet insist that they can see perfectly and hemineglect patients will bluntly assert that a paralysed arm is functionally normal (see Ramachandran and Blakeslee (1998) for examples). In the face of these cases, it cannot be simply assumed that it would be impossible for us to be systematically mistaken about our phenomenal states. Further criticisms of Chalmers' argument can be found in Van Heuveln et. al (1998) and Prinz (2003).

3.3 Correlates and causes of consciousness

It might be objected that a great deal of progress has been made with identifying the correlates of consciousness, which in the longer term may enable us to work out what its causes are. Crick and Koch (2003) give a nice overview of this work and a more specific example would be Aleksander's (2000) connection between gaze-locked cells (identified by Galletti and Battaglini (1989)) and our experience of stable objective space. Eventually external observers may be able to use a brain scan to make a precise description of a person's conscious mental content.

However, whilst this work shows that certain patterns of firing neurons or synchronization between them are necessary and perhaps sufficient correlates of consciousness *in the human case*, they do not show that these are necessary and sufficient correlates of consciousness *in general*. All of the experiments on the correlates of consciousness have been carried out on biological subjects and so it is not clear whether the brain's functions are correlated with consciousness by themselves or whether a biological substance is also necessary. Without systematic separation of the factors it is impossible to say whether a robot with the same global functions as a human would experience phenomenal states.

Taken together, the arguments in this section force us to the conclusion that no test can separate out necessary and sufficient correlates or causes of consciousness. We can vary the ways in which the global functions of the brain are implemented in a vast number of ways, but since these will always lead to the same behavioural output, any impact of these changes on consciousness cannot be measured and we will never know for certain whether a functionally (and thus behaviourally) identical robot has conscious states or not.

4 Ordinal Probability Scale

Faced with these difficulties we could follow Prinz (2003) and suspend judgement about whether robots built from different principles are capable of supporting phenomenal states. However, there are three problems with this mysterianism. To begin with, we have a strong intuition that machines built along similar lines to human beings are likely to be phenomenally conscious. The more similar a system is to human beings, the more likely we are to believe that it experiences conscious states of some kind. Second, as machine consciousness develops we will be developing machines that exhibit increasingly complex behaviour and spend a lot of time in confused states and potentially in pain. This has been somewhat dramatically compared by Metzinger (2003) to the development of a race of retarded infants for experimentation. To address these ethical worries without stifling research a way needs to be found to evaluate the probability that a robot is experiencing phenomenal states. A third problem with mysterianism is that as more sophisticated robots emerge, people are inevitably going to attribute more and more conscious states to them. People already attribute feelings to Kismet or AIBO, and a systematic way of evaluating phenomenal probability needs to be in place before this becomes a live public issue. The general public is very interested in the question whether something

is *really* conscious and it would be helpful if the machine consciousness community could formulate some kind of reply, even if this is based on analogy with human beings.

To address these issues and provide a framework within which the more detailed work of synthetic phenomenology can proceed, I propose the construction of a probability scale that orders machines according to the probability that their architecture is capable of supporting conscious states. This says nothing about whether a machine is *at present* conscious (this is the task of the second and third stages of synthetic phenomenology outlined in section 2); only whether it is likely that this kind of system can support conscious states.

I will start this description of the scale with an overview of the systems that are covered by it. After explaining the factors and the way in which they are combined, I will give a few specific examples to illustrate how it works.

4.1 Systems covered by this scale

This scale only covers systems that approximate the *global* functions of a human brain. By global functions I mean the functions that transform the brain's sensory inputs into motor outputs along the nerves connecting the brain to the body. Such a system could either be used to control a real human body or it could have its own real or virtual artificial body. In the latter case, the artificial body would have to have approximately the same number and type of sensors and effectors as the human body with approximately the same resolution.

The notion of approximating the global functions of the human brain is defined here using Harnad's (1994) extended T3 version of the Turing test. A machine that approximates the functions of a human brain by controlling a human or artificial body would have to be completely indistinguishable in external function from humans for 70 years or more. Such a robot could hold down a job, create works of art and have relationships with other human beings. Machines that were interned in an asylum for strange behaviour would not be considered functionally identical to a human being.

4.2 Factors affecting the probability of phenomenal consciousness

This scale is constructed in relation to humans, who are at present the benchmark example of conscious machines. The more similar a machine is to a human, the more likely it is to be phenomenally conscious. The factors within each group are assigned weightings (W) ranging from 1.0 to 0.1. These are arbitrary values and the way in which they are combined and converted into an ordinal scale is explained in section 4.3. An outline of the factors that I have selected for this first draft of the probability scale now follows.

4.2.1 Rate

Machines can operate much faster or slower than the human brain and we are more likely to attribute consciousness to a machine that runs at approximately the same speed. If we were forced to say whether the economy of Bolivia or the Earth's crust is more likely to be conscious, we would probably choose the economy of Bolivia. This is not because it is more complex or has more states, but because its states change more rapidly.

Table 1: Rate factors

	Rate	**W**
R1	Approximately the same speed as human brain	1.0
R2	10 times faster or slower than human brain	0.55
R3	Over 100 times faster or slower than human brain	0.1

4.2.2 Size

We are more likely to attribute consciousness to a system that fits inside a person's head, than to a system that is the size of the population of China.

Table 2: Size factors

	Size	**W**
S1	Approximately the same size as human brain	1.0
S2	1000 times larger or smaller than human brain	0.55
S3	More than a million times larger or smaller than human brain	0.1

4.2.3 Functional granularity

This probability scale keeps the global functions of the brain constant. However, there is a wide variety of ways in which the global functions of the brain can be implemented by different collections of local functions, some of which are closer to the human brain than others. This factor weights machines according to the degree to which their functional granularity matches that of the human brain. I have gone down to the atomic level to take account of claims by Hameroff and Penrose (1996) that consciousness depends on quantum functions.

This factor is complicated by the fact that neurons can be used to implement functions in a biological and non-biological way. For example, the function of the whole brain could be implemented

by an vast neural network trained by back propagation, or it could be implemented by a more biological structure of neurons. Since neurons can themselves be simulated using neurons there is also potentially infinite self-recursion, which I have limited by a restriction introduced in section 4.3. To keep things simple I have set aside the possibility that glia play an information-processing role.

The way in which these four tables are combined is fairly self-evident. If the brain's global functions are implemented by a biological structure of modules, then the way in which the functions of the modules are implemented has to be specified as well. On the other hand, no further levels are required if the brain's global functions are implemented by a simulation that is not biologically structured.

Table 3: Whole brain function

	Function of whole brain	**W**
FW1	Produced by a biological structure of modules	1.0
FW2	Produced by a non-biological structure of modules	0.7
FW3	Produced by a non-biological structure of neurons	0.4
FW4	Simulated using mathematical algorithms, computer code or some other method	0.1

Table 4: Module functions

	Function of modules	**W**
FM1	Produced by a biological structure of neurons	1.0
FM2	Produced by a non-biological structure of neurons	0.7
FM3	Produced by a mixture of methods	0.4
FM4	Simulated using mathematical algorithms, computer code or some other method	0.1

Table 5: Neuron function

	Function of neurons	**W**
FN1	Produced by a biological structure of molecules, atoms and ions	1.0
FN2	Produced by a non-biological structure of molecules, atoms and ions (silicon chemistry, for example)	0.7
FN3	Produced by a non-biological structure of neurons	0.4
FN4	Simulated using mathematical algorithms, computer code or some other method	0.1

Table 6: Function of molecules, atoms and ions

	Function of molecules, atoms and ions	**W**
FMAI1	Produced by real subatomic phenomena, such as protons, neutrons and electrons	1.0
FMAI2	Produced by a non-biological structure of neurons	0.55
FMAI3	Simulated using mathematical algorithms, computer code or some other method	0.1

4.2.4 Simulation time-slicing

The simulation of brain functions can be carried out in parallel with all the different functions working simultaneously on dedicated hardware. On the other hand a single processor can emulate the parallel operation of many functions by time-slicing. This scale follows Kent (1981) in ranking time-sliced simulations, which only have the same time complexity as the brain, as being less likely to be phenomenally conscious than simulations whose parts have the same moment-to-moment space complexity as the brain. In this first draft of this scale, I have placed all of the different types of simulation hardware together – such as a modern computer built from silicon and copper, a light computer, Searle's Chinese room or the economy of Bolivia. I have also set aside the potential question about the link between consciousness and virtual machines.

Table 7: Simulation time slicing

	Simulation time slicing	**W**
STS1	Complete hardware simulation in which all parts of the model are dynamically changing and co-present at any point in time	1.0
STS2	Multi-processor time-sliced simulation in which only parts of the model are dynamically changing and co-present at any point in time	0.55
STS3	Single processor time-sliced simulation in which only a single part of the model is dynamically changing and present at any point in time	0.1

4.2.5 Analogue / digital

With an analogue simulation there is an infinity of possible states, which can only be approximated by a digital simulation. It is possible that some nonlinear properties of the brain are more faithfully captured by an analogue simulation.

Table 8: Analogue / digital simulation

	Analogue / digital	**W**
AD1	Analogue simulation	1.0
AD2	Mixture of analogue and digital	0.55
AD3	Digital simulation	0.1

4.3 Putting it all together

To obtain the final ordinal probability scale, a complete list of all the possible machines is extracted from the factor tables. The weightings that are applicable to each machine are then multiplied together to give a total weighting for each machine. These are then used to situate all of the different machines in an ordinal scale. Since many of the machines have the same total weighting, this scale is much shorter than the total number of possible combinations. I have also had to introduce a couple of extra rules for the combination of factors:

1. Since neurons can be used to simulate the behaviour of neurons, or the molecules/atoms/ions that neurons are composed of, the functional granularity is potentially infinitely self-recursive. To prevent this I have stipulated that if non-biological structures of neurons are used to implement the functions of neurons or molecules/atoms/ions, then the neurons that are used for this cannot themselves have their functions implemented using non-biological structures of neurons.
2. When machines have less functional granularity than the brain some kind of penalty needs to be imposed on machines that deviate from the human structure – for example, when the function of whole brain is implemented by a complex lookup table. In the present implementation, the number of levels of a human brain is 4 and so I will use 0.1 as the weighting for each missing level of functional granularity.

This scale starts with human beings and finishes with digital single-processor simulations based on non-biological principles that are much larger or smaller than the human brain and process at a much slower or faster rate. There is not space in this paper to list all the possible combinations of factors in a single ordinal scale – the complete list has over a million combinations. Instead, I have integrated everything together on a webpage,[2] which can be used to calculate the position of a machine on the scale. Some examples are given in the next section.

4.4 Examples

None of the systems discussed in this section are even close to reproducing the global functions of the human brain. However, to illustrate how this scale could work in practice, I will assume that these examples have developed to the point at which they could pass the T3 version of the Turing test.

4.3.1 Neurally Controlled Animat

This is a system developed by DeMarse et al. (2001) that uses biological neurons to control a simulated body in a virtual world. The biological neurons are initially disassociated and then self-assemble in response to stimulation from their environment. Since the organisation of the neurons is not determined by the many factors present in embryological development, this system produces the functions of the whole brain from a non-biological structure of neurons. The factors are: R1, S1 FW3, FN1 and FMAI1, giving a total weighting of 0.4, This needs to be multiplied by 0.1 to compensate for the lack of functional granularity at the level of modules and so the total weighting is 0.04, which works out as an ordinal ranking of 48 out of 812.

4.3.2 Lucy

Lucy is a robot developed by Grand (2003) that is controlled by a multi-processor simulation of neurons arranged into a biological structure. The factors are thus R1, S1, FW1, FM1, FN4, STS2 and AD3 giving a total weighting of 5.5×10^{-3}. This needs to be multiplied by 0.1 to compensate for the lack of functional granularity at the level of molecules, atoms and ions, and so the total weighting becomes 5.5×10^{-4}. This gives Lucy an ordinal ranking of 285 out of 812.

4.3.3 IDA

IDA is a naval dispatching system created by Franklin et. al. (1998). This system is based on Baars (1988) global workspace model of consciousness and so its modules could be said to be biologically structured. However the solutions that are used to implement the different modules are non-biological. The factors are R1, S1, FW1, FM4, STS2 and AD3. This gives a total weighting of 5.5×10^{-3}, but since the functional granularity is less than

[2] http://www.syntheticphenomenology.net

the human brain by two levels, this weighting needs to be multiplied by 0.01, to give a total weighting of 5.5×10^{-5}, which is an ordinal ranking of 461 out of 812.

4.3.4 The population of China

This is a thought experiment suggested by Block (1978) in which the functions of a human brain are carried out by the population of China interconnected by two-way radios and satellites. This is a non-biological structure in which modules assembled from biological neurons are combined with modules built with other hardware. The population of China is approximately 1.3 billion and so this 'machine' is very much larger than the human brain. It is also likely to work at a much slower rate. One problem with Block's thought experiment is that the details about the functional implementation are left very vague and so I have classified it as multi-processor hardware combined with modules assembled from neurons simulated using biological neurons. The factors are: R3, S3, FW2, FM3, MSTS2, MAD3, FN3, FNN1 and FMA1, which gives a total weighting of 6.16×10^{-5}. This works out as an ordinal ranking of 445 out of 812. Although this seems surprisingly high, it is the presence of biological hardware (organised in a non-biological way) that elevates it above systems that are purely based on simulation. 1.3 billion computers networked together to produce the same result would have a ranking of 786 out of 812.

5 Discussion

A number of issues arise in connection with this probability scale:

1) To begin with, it is at present unclear whether consciousness decreases gradually as we move away from the human machine, or whether there is a cut off point at which consciousness simply vanishes. Consciousness may simply cease to exist in a system unless neurons are simulated at the molecular level, or a cluster of factors may interact in a critical way such that phenomenal states cannot be produced without one of the factors. If consciousness cuts off abruptly, then this ordinal probability scale expresses the likelihood that consciousness is present in a machine built in a particular way. On the other hand, if consciousness decreases gradually as the factors become less human, then this ordinal scale ranks machines according to their level of consciousness.

2) This is an extremely anthropocentric probability scale. The great chain of machines is a kind of fall from grace from perfectly conscious man. This is an epistemological necessity – we only know for sure that we are conscious – but it is quite possible, although empirically undeterminable, that robots at the far end of the probability scale are more conscious than ourselves. This scale is a probabilistic rating based on our guess that machines built along lines similar to our own (such as other people) are more likely to experience phenomenal states than machines built along lines very different from our own. I believe that such a scale could be useful, but it should not be taken as anything more than the systematisation of an intuition.

3) This scale does not explicitly list many of the factors that have been put forward as potential correlates of consciousness. However, many of these are implicit possibilities within the available architectures. For example, re-entrant connections are assumed to be possible within any of the machines that have biologically structured neurons. However, there will inevitably be some factors that are not included within this version of the scale, which can be added to subsequent versions.

4) This scale only applies to machines whose global functions approximate those of the human brain. A perpendicular scale could be added that orders people, machines and animals according to the *degree* to which their global functions approximate those of the human brain. Machines with functions processing visual data about faces might be ranked higher on this scale than machines that analyse banking details. The more the system's global functions match those of the human brain, the more likely it would be to possess phenomenal consciousness (or the more phenomenal consciousness it would possess).

5) It is worth noting that I have set aside the whole question of the body here. In theory a computer could approximate the global input and output functions of the brain without inhabiting a body at all. However, such a system would be almost impossible to develop and, according to Damasio (1995), there may be a critical link between the body and consciousness.

6) Finally, this scale is likely to become superfluous when we eventually achieve machine consciousness. When we talk to robots every day, work with robots that display conscious behaviour and perhaps even marry robots with emotional functions, we will cease to worry about whether they *really* have phenomenal states; just as we rarely think that other people are automatons.

6 Previous Work

Synthetic phenomenology is an area that is only just starting to receive detailed attention. According to Chrisley (2004), the term first made its appearance in the machine consciousness community at the Models of Consciousness Workshop (held in Birmingham 2003) and was independently coined by Scott Jordan (1998). It is related to *synthetic epistemology*, which is defined by Chrisley and Holland (1994, p. 1) as the "creation and analysis of artificial systems in order to clarify philosophical issues that arise in the explanation of how agents, both natural and artificial, represent the world." Since this area is so new, relatively little research has actually been carried out on it. The work that has been done includes Chrisley's (1995) analysis of non-conceptual content and Holland and Goodman's (2003) use of inversion to map out a robot's internal representations.

The question about phenomenal states in robots has been extensively discussed in the literature on consciousness. The contributions roughly divide into those who accept the difficulties with behaviour-based attribution of phenomenal states, and those with a theory of consciousness that enables them to make definite claims about which machines are phenomenally conscious. In the first group, Moor (1988) sets out the arguments against knowing for certain whether robots have qualia, but claims that we will need to attribute qualia to robots in order to understand their actions. A similar position is set out by Harnad (2003), who accepts the behaviour-based arguments set out by Moor and Prinz, but claims that the other minds problem means that we can only ever attribute consciousness on the basis of behaviour and so any robot that passes the T3 version of the Turing test for a lifetime must be acknowledged to be conscious. Prinz (2003) is closest to the position of this paper since he does not think that we can identify the necessary and sufficient conditions for consciousness and does not suggest other grounds for attributing consciousness to machines.

People who claim to know exactly what the causes or correlates of consciousness are can say precisely which machines are capable of phenomenal states; replacing the ordinal probability scale set out in this paper with a dividing line dictated by their theory of consciousness. One of the most liberal of these theories is Chalmers (1996), whose link between consciousness and information leads him to attribute limited phenomenal states to machines as simple as thermostats. At the other extreme, Searle (1980) believes that his Chinese room argument excludes the possibility that any of the levels of functional granularity could be simulated and rather vaguely ties consciousness to a causal property of matter, so that only biological humans, animals and possibly aliens could be conscious. In between these positions are people like Aleksander and Dunmall (2003), who suggests five necessary conditions or axioms for consciousness. According to Aleksander and Dunmall, machines can only be conscious if they have depiction, imagination, attention, planning and emotion.

7 Conclusion

In this paper I have set out a proposal for an ordinal probability scale, which can be used to assess the likelihood that a machine is capable of experiencing phenomenal states. This scale only applies to machines that can pass the T3 version of the Turing test by controlling a human or artificial body. This scale can help us to evaluate the ethical significance of machine consciousness experiments and in some cases it could be used to select a machine implementation that has less probability of phenomenal suffering. The scale put forward in this paper is only a first draft with some of the factors that may be correlates of consciousness. If it is found useful, I hope that it will be improved by other people and perhaps develop into a standard as we get closer to realising conscious machines.

Acknowledgements

Many thanks to Owen Holland for feedback and comments about this paper. Thank you also to the EPSRC for funding this project.

References

Igor Aleksander and Barry Dunmall, An extension to the hypothesis of the asynchrony of visual consciousness. *Proceedings of the Royal Society of London B*, 267: 197-200, 2000.

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.

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

Ned Block. Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science*, Volume IX, *Perception and Cognition Issues in the Foundations of Psychology*, edited by C. Wade

Savage, Minneapolis: University of Minnesota Press, 1978.

David Chalmers. *The Conscious Mind*. Oxford: Oxford University Press, 1996.

Ronald J. Chrisley and Andy Holland. Connectionist Synthetic Epistemology: Requirements for the Development of Objectivity. *COGS CSRP*, 353: 1-21, 1994.

Ronald J. Chrisley. Taking Embodiment Seriously: Nonconceptual 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.

Ronald J. Chrisley. Synthetic Phenomenology. Talk at the Workshop on Machine Consciousness, Antwerp, 28th June 2004.

Francis Crick and Christof Koch. A framework for consciousness. *Nature Neuroscience*, 6(2): 119-26, 2003.

A. R. Damasio. *Descartes' Error: emotion, reason and the human brain*. London : Picador, 1995.

S. Dehaene and L. Naccache. Towards a cognitive neuroscience of consciousness : Basic evidence and a workspace framework. *Cognition*, 79: 1-37, 2001.

T. B. DeMarse, D. A. Wagenaar, A. W. Blau, and S. M. Potter. The Neurally Controlled Animat: Biological Brains Acting With Simulated Bodies. *Autonomous Robots,* 11(3): 305-310, 2001.

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

Claudio Galletti and Piero Paolo Battaglini. Gaze-Dependent Visual Neurons in Area V3A of Monkey Prestriate Cortex. *The Journal of Neuroscience,* 9(4): 1112-1125, 1989.

Steve Grand. *Growing up with Lucy*. London: Weidenfeld & Nicolson, 2003.

Stuart Hameroff, and Roger Penrose. Orchestrated Reduction Of Quantum Coherence In Brain Microtubules: A Model For Consciousness? In S.R. Hameroff, , A.W. Kaszniak, and A.C. Scott (eds), *Toward a Science of Consciousness - The First Tucson Discussions and Debates*, Cambridge, MA: MIT Press, 507-540, 1996.

Stevan Harnad. Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life. *Artificial Life* 1(3): 1994.

Stevan Harnad. Can a Machine Be Conscious? How? In Owen Holland (ed.), *Machine Consciousness*, Exeter: Imprint Academic, 2003.

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

J. S. Jordan. Synthetic phenomenology? Perhaps, but not via information processing. Talk given at the Max Planck Institute for Psychological Research, Munich, Germany, 1998.

Ernest W. Kent. *The Brains of Men and Machines*. Peterborough: BYTE/ McGraw Hill, 1981.

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

J.H. Moor. Testing robots for qualia. In H.R. Otto and J.A. Tuedio (eds), *Perspectives on Mind*, Dordrecht/ Boston/ Lancaster/ Tokyo: D. Reidel Publishing Company, 1988.

Jesse J. Prinz. Level-Headed Mysterianism and Artificial Experience. In Owen Holland (ed.), *Machine Consciousness*, Exeter: Imprint Academic, 2003.

V. S. Ramachandran and S. Blakeslee. *Phantoms in the Brain*. London: Fourth Estate, 1998.

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

J. Searle. *The Rediscovery of the Mind*. Cambridge, Massachusetts: MIT Press, 1992.

B. Van Heuveln, E. Dietrich and M. Oshima. Let's dance! The equivocation in Chalmers' dancing qualia argument. *Minds and Machines,* 8: 237-49, 1998.

S. Zeki and A. Bartels. The asynchrony of consciousness. *Proceedings of the Royal Society B*, 265: 1583-1585, 1998.

J. Zihl, D. Von Cramon and N. Mai. Selective Disturbance of Movement Vision after Bilateral Brain Damage. *Brain,* 106: 313-340, 1983.

Simulation and Representation of Body, Emotion, and Core Consciousness

Tibor Bosse[*] Catholijn M. Jonker[†] Jan Treur[*]

[*]Vrije Universiteit Amsterdam
Department of Artificial Intelligence
De Boelelaan 1081a
1081 HV Amsterdam
The Netherlands
{tbosse, treur}@cs.vu.nl

[†]Radboud Universiteit Nijmegen
Nijmegen Institute for Cognition and Information
Montessorilaan 3
6525 HR Nijmegen
The Netherlands
C.Jonker@nici.ru.nl

Abstract

This paper contributes an analysis and formalisation of Damasio's theory on core consciousness. Three important concepts in this theory are "emotion", "feeling", and "feeling a feeling" (or core consciousness). In particular, a simulation model is described of the neural dynamics leading via emotion and feeling to core consciousness, and dynamic properties are formally specified that hold for these dynamics. These properties have been automatically checked for the simulation traces. Moreover, a formal analysis is made and verified of relevant notions of representation.

1 Introduction

In (Damasio, 2000) the neurologist Antonio Damasio puts forward his theory of consciousness. He describes his theory in an informal manner, and supports it by a vast amount of evidence from neurological practice. More experimental work supporting his theory is reported in (Damasio et al., 2000; Parvizi and Damasio, 2001). Damasio's theory is described on the one hand in terms of the occurrence of certain neural states (or neural patterns), and temporal or causal relationships between them. Formalisation of these relationships requires a modelling format that is able to express direct temporal or causal dependencies. On the other hand Damasio gives interpretations of most of these neural states as representations, for example as 'sensory representation', or 'second-order representation'. This requires an analysis of what it means that a neural state is a representation for something. This paper focuses on Damasio's notions of 'emotion', 'feeling', and 'core consciousness' or 'feeling a feeling'. In (Damasio, 2000), Damasio describes an *emotion as neural object* (or *internal emotional state*) as an (unconscious) neural reaction to a certain stimulus, realized by a complex ensemble of neural activations in the brain. As the neural activations involved often are preparations for (body) actions, as a consequence of an internal emotional state, the body will be modified into an *externally observable emotional state*. Next, a *feeling* is described as the (still unconscious) sensing of this body state. Finally, *core consciousness* or *feeling* a *feeling* is what emerges when the organism detects that its representation of its own body state (the *proto-self*) has been changed by the occurrence of the stimulus: it becomes (consciously) aware of the feeling.

This paper aims at formalisations and simulation models for these three notions. In addition, the notion of representation used by Damasio is formally analysed against different approaches to representational content from the literature on the Philosophy of Mind. It is shown that the classical causal/correlational approach to representational content, e.g., (Kim, 1996), pp. 191-193, is inappropriate to describe the notion of representation for core consciousness used by Damasio, as this notion essentially involves more complex temporal relationships describing histories of the organism's interaction with the world. An alternative approach is shown to be better suited: representational content as relational specification over time and space, cf. (Kim, 1996), pp. 200-202. Criteria for this approach are formalised, and it is shown that the formalisation of Damasio's notions indeed fit these criteria.

A brief summary of the main basic assumptions underlying Damasio's approach is expressed in: 'First, I am suggesting that (…) 'having a feeling' is not the same as 'knowing a feeling', that reflection on feeling is yet another step up. (…) The inescapable and remarkable fact about these three phenomena – emotion, feeling, conscious-

ness – is their body relatedness. (…) As the representations of the body grow in complexity and coordination, they come to constitute an integrated representation of the organism, a proto-self. Once that happens, it becomes possible to engender representations of the proto-self as it is affected by interactions with a given environment. It is only then that consciousness begins, only thereafter that an organism that is responding beautifully to its environment begins to discover that *it* is responding beautifully to its environment. But all of these processes – emotion, feeling, and consciousness – depend for their execution on representations of the organism. Their shared essence is the body. (Damasio, 2000), pp. 283-284.

In Section 2 the modelling approach used is briefly introduced. In Sections 3, 4, and 5, for a simple example models are presented for the processes leading to emotion, feeling, and feeling a feeling (or conscious feeling), respectively. Section 6 provides the results of a simulation of these models. In Section 7 it is analysed in how far the representational content of Damasio's notions can be described by two approaches from Philosophy of Mind. Formalisations of some of the dynamic properties of the processes leading to emotion, feeling and feeling a feeling are presented. Next, Section 8 addresses verification. It is shown that the notions for representational content developed in Section 7 indeed hold for the model. The verification is performed both by automated checks and by mathematical proof. Section 9 concludes the paper with a discussion.

2 Modelling Approach

To model the making of emotion, feeling and core consciousness, dynamics play in important role. Dynamics will be described in the next section as evolution of *states* over time. The notion of state as used here is characterised on the basis of an ontology defining a set of state properties that do or do not hold at a certain point in time. The modelling perspective taken is not a symbolic perspective, but essentially addresses the neural processes and their dynamics as neurological processes. This implies that states are just neurological states. To successfully model such complex processes, forms of abstraction are required; for example:

- neural states or activation patterns are modelled as single state properties
- large-dimensional vectors of such (distributed) state properties are composed to one single composite state property, when appropriate; e.g., (p1, p2, …) to p and (S1, S2, …) to S in Section 3.

To describe the dynamics of the processes mentioned above, explicit reference is made to time. Dynamic properties can be formulated that relate a state at one point in time to a state at another point in time. A simple example is the following dynamic property specification for belief creation based on observation:

'at any point in time t1, if the agent observes rain at t1,
then there exists a point in time t2 after t1 such that
at t2 the agent has internal state property s'

Here, for example, s can be viewed as a sensory representation of the rain. To express dynamic properties in a precise manner a language is used in which explicit references can be made to time points and traces: the Temporal Trace Language TTL; cf. (Jonker and Treur, 2002). Here a *trace or trajectory* over an ontology Ont is a time-indexed sequence of states over Ont. The sorted predicate logic temporal trace language TTL is built on atoms referring to, e.g., traces, time and state properties. For example, 'in the internal state of agent A in trace γ at time t property s holds' is formalised by state(γ, t, internal(A)) |= s. Here |= is a predicate symbol in the language, usually used in infix notation, which is comparable to the Holds-predicate in situation calculus. Dynamic properties are expressed by temporal statements built using the usual logical connectives and quantification (for example, over traces, time and state properties).

To be able to perform some (pseudo)-experiments, a simpler temporal language has been used to specify simulation models in a declarative manner. This language (the *leads to* language) enables to model direct temporal dependencies between two state properties in successive states. This executable format is defined as follows. Let α and β be state properties of the form 'conjunction of atoms or negations of atoms', and e, f, g, h, non-negative real numbers. In the *leads to* language the notation $\alpha \rightarrow\!\!\!\!\rightarrow_{e, f, g, h} \beta$, means:

If state property α hold for a time interval with duration g,
then after some delay (between e and f) state property β will hold
for a time interval of length h.

For a precise definition of the *leads to* format in terms of the language TTL, see (Jonker, Treur, and Wijngaards, 2003). A specification of dynamic properties in *leads to* format has as advantages that it is executable and that it can often easily be depicted graphically.

In Sections 3, 4, 5 and 6, the *leads to* format has been used to create simulation models of the processes leading to emotion, feeling and core consciousness in terms of neural processes. Given this physical-level model and its dynamic properties, a next step is to assign representational content to (some of) the relevant state properties. For nontrivial cases representational content involves histories of interaction between organism and world (Bickhard, 1993; Jonker and Treur, 2003), and this also

shows up in Damasio's theory. To specify and analyse the representational content to a number of state properties of the models and the traces they generate, the more expressive TTL format is used in Section 7. Both formats are used in Section 8.

3 Emotion

First Damasio's notion of *emotion* is addressed. He explains this notion as follows: 'The substrate for the representation of emotions is a collection of neural dispositions in a number of brain regions (…) They exist, rather, as potential patterns of activity arising within neuron ensembles. Once these dispositions are activated, a number of consequences ensue. On the one hand, the pattern of activation represents, within the brain, a particular emotion as 'neural object'. On the other, the pattern generates explicit responses that modify both the state of the body proper and the state of other brain regions. By so doing, the responses create an emotional state, and at that point, an external observer can appreciate the emotional engagement of the organism being observed. (Damasio, 2000), p. 79. According to this description, an *internal emotional state* is a collection of neural dispositions in the brain, which are activated as a reaction on a certain stimulus. Once such an internal emotional state occurs, it entails modification of both the body state and the state of other brain regions. By these events, an *external emotional state* is created, which is accessible for external observation.

Assume that the music you hear is so special that it leads to an emotional state in which you show some body responses on it (e.g., shivers on your back). This process is described by executable local dynamic properties taking into account internal state properties sr(music) for activated sensory representation of hearing the music, and (p1, p2, …) a vector for the activation of preparatory states for the body responses (S1, S2, …); see Figure 1.

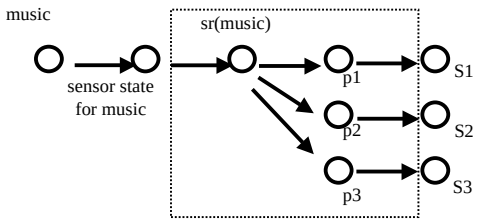


Figure 1: Processes leading to an emotional state

These vectors are the possible internal emotional states. Note that the state properties are abstract in the sense that a state property refers to a specific neural activation pattern. In the model the conjunction p1 & p2 & .. of these preparatory state properties is denoted by p; this p can be considered a composite state property. Moreover, the conjunction of the vector of all body state properties responding to the music S1, S2, … (i.e., the respective body state properties for which p1, p2, ... are preparing) is denoted by (composite) state property S.

The model abstracted in this manner is depicted in Figure 2, upper part. In formal textual format these local properties are as follows:

LP0 music •→» sensor_state(music)
LP1 sensor_state(music) •→» sr(music)
LP2 sr(music) •→» p
LP3 p •→» S

In the remainder of this paper this abstract type of modelling will be used. Notice, however, that each of the abstract state properties used are realised in the organism in a distributed manner as a large-dimensional vector of more local (neural) state properties. Also the sensory representation sr(music) may be considered such a composite state property with different aspects of the music represented in different forms at different places. Notice, moreover, that the names of the state properties have been chosen to support readability for humans. But in principle these names should be considered as neutral indications of neural states, such as n1, n2, and so on.

4 Feeling

Next, Damasio's notion of *feeling* is considered. He expresses the emergence of feeling as follows: As for the internal state of the organism in which the emotion is taking place, it has available both the emotion as neural object (the activation pattern at the induction sites) and the sensing of the consequences of the activation, a feeling, provided the resulting collection of neural patterns becomes images in mind. (…) The changes related to body state are achieved by one of two mechanisms. One involves what I call the 'body loop'. (…) .. the body landscape is changed and is subsequently represented in somatosensory structures of the central nervous system, from the brain stem on up. The change in the representation of the body landscape can partly be achieved by another mechanism, which I call the 'as if body loop'. In this alternate mechanism, the representation of body-related changes is created directly in sensory body maps, under the control of other neural sites, for instance, the prefrontal cortices. It is 'as if' the body had really been changed but it was not. (…) Assuming that all the proper structures are in place, the processes reviewed above allow an organism to undergo an emotion, exhibit it, and image it, that is, feel the emotion. (Damasio, 2000), pp. 79-80. Thus, a feeling emerges when the collection of neural patterns contributing to the emotion lead to mental images. In other words, the organism senses the consequences of the internal emotional state. Damasio distinguishes two mechanisms by which a feeling can be achieved:

1) Via the *body loop*, the internal emotional state leads to a changed state of the body, which subsequently, after sensing, is represented in

somatosensory structures of the central nervous system.

2) Via the *as if body loop*, the state of the body is not changed. Instead, on the basis of the internal emotional state, a changed representation of the body is created directly in sensory body maps. Consequently, the organism experiences the same feeling as via the body loop: it is 'as if' the body had really been changed but it was not.

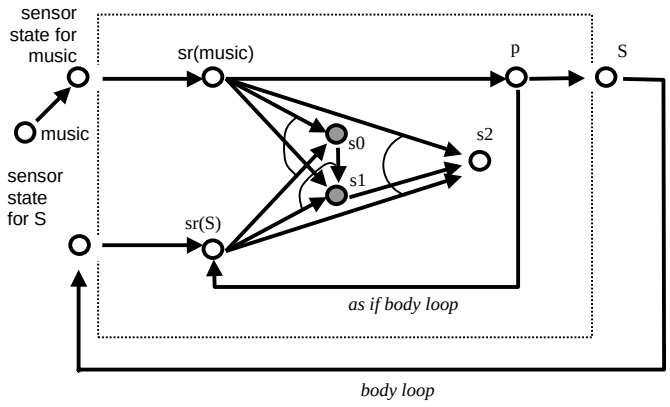


Figure 2: Overview of the simulation model

The model described in Section 3 can be extended to include a number of internal state properties for sensory representations of body state properties that are changed due to responses on the music; together these sensory representations constitute the feeling induced by the music. In Figure 2 the conjunction of these sensory representations is depicted: sr(S) (a sensory representation of the changed body state; this may be materialised in a distributed manner as a kind of vector). This describes the 'body loop' for the responses on the music; here S and sensor_state(S) are effects and sensors in the body, respectively. In formal format, two additional local dynamic properties are needed (see also Figure 2):

LP4 S $\bullet\!\!\rightarrow\!\!\!\!\rightarrow$ sensor_state(S)
LP5 sensor_state(S) $\bullet\!\!\rightarrow\!\!\!\!\rightarrow$ sr(S)

Notice that an internal state property sr(shivering) for shivering only, does not directly relate to the music. It is caused by the external stimulus shivering, which in this particular case is originally caused by the music. This body state property shivering could be present for a lot of other reasons as well, e.g., a cold shower. However, taking into account that not only shivering but a larger number of sensory state properties constitute the overall composite state property sr(S), the feeling will be more unique for the music. For the case of an 'as if body loop' dynamic properties LP3, LP4 and LP5 can be replaced by the following local dynamic property directly connecting p and sr(S).

LP6 p $\bullet\!\!\rightarrow\!\!\!\!\rightarrow$ sr(S)

Also a combination of models can be made, in which some effects of hearing the music is caused by a body loop and some are caused by an 'as if body loop'.

5 Feeling a Feeling

Finally, Damasio's notion of *knowing* or *being conscious of* or *feeling* a *feeling* is addressed. This notion is based on the organism detecting that its representation of its own (body) state (the *proto-self*) has been changed by the occurrence of a certain object (the music in our example). According to Damasio, the proto-self is "a coherent collection of neural patterns which map, moment by moment, the state of the physical structure of the organism". (Damasio, 2000), p. 177. He expresses the way in which the proto-self contributes to a conscious feeling in the following hypothesis: Core consciousness occurs when the brain's representation devices generate an imaged, nonverbal account of how the organism's own state is affected by the organism's processing of an object, and when this process enhances the image of the causative object, thus placing it in a spatial and temporal context. (p. 169)… with the license of metaphor, one might say that the swift, second-order nonverbal account narrates a story: *that of the organism caught in the act of representing its own changed state as it goes about representing something else*. But the astonishing fact is that the knowable entity of the catcher has just been created in the narrative of the catching process. (…) You know it is *you* seeing because the story depicts a character – you – doing the seeing. (pp. 170-172) … beyond the many neural structures in which the causative object and the proto-self changes are separately represented, there is at least one other structure which *re-represents* both proto-self and object in their temporal relationship and thus represent what is actually happening to the organism: *proto-self at the inaugural instant; object coming into sensory representation; changing of inaugural proto-self into proto-self modified by object*. (p. 177; italics in the original). In summary, the conscious feeling occurs when the organism detects the transitions between the following moments:

1. The proto-self exists at the inaugural instant.
2. An object comes into sensory representation.
3. The proto-self has become modified by the object.

For our case we restrict ourselves to placing the relevant events in a temporal context. In a detailed account, in the trace considered subsequently the following events take place: no sensory representations for music and S occur, the music is sensed, the sensory representation sr(music) is generated, the preparation representation p for S is generated, S occurs, S is sensed, the sensory representation sr(S)

is generated. According to Damasio (2000), pp. 177-183, two transitions are relevant (see Damasio's Figure 6.1), and have to be taken into account in a model:

- from the sensory representation of the initial no S body state and not hearing the music to hearing music and a sensory representation of the music, and no S sensory representation
- from a sensory representation of the music and no sensory representation of S to a sensory representation of S and a sensory representation of the music

These two transitions are to be detected and represented by the organism. To model this process three internal state properties are introduced: s0 for encoding the initial situation, and s1 and s2 subsequently for encoding the situations after the two relevant changes. By making these state properties persistent they play the role of indicating that in the past a certain situation has occurred. Local dynamic properties that relate these additional internal state properties to the others can be expressed as follows (see also Figure 2):

LP7 not sr(music) & not sr(S) •→→ s0
LP8 sr(music) & not sr(S) & s0 •→→ s1
LP9 sr(music) & sr(S) & s1 •→→ s2

State properties s0 and s1 are persistent.

6 Simulation

A special software environment has been created to enable the simulation of executable models (Bosse et al., 2004). Based on an input consisting of dynamic properties in *leads to* format (and their timing parameters e, f, g, h, see Section 2), this software environment generates simulation traces. The algorithm used for the simulation is rather straightforward: at each time point, a bound part of the past of the trace (the maximum of all g values of all rules) determines the values of a bound range of the future trace (the maximum of f + h over all LEADSTO rules). The software was written in SWI-Prolog/XPCE, and consists of approximately 20000 lines of code. For more implementation details, see (Bosse et al., 2004).

Using this software environment, the model described in the previous sections has been used to generate a number of simulation traces. An example of such a simulation trace can be seen in Figure 3. Here, time is on the horizontal axis, the state properties are on the vertical axis. A dark box on top of the line indicates that the property is true during that time period, and a lighter box below the line indicates that the property is false. This trace is based on all executable local properties (i.e., LP0 to LP9), except LP6. In all properties, the values (0,0,1,1) have been chosen for the timing parameters e, f, g, and h. Figure 3 shows how the presence of the music first leads to an emotion (p or S), then to a feeling (sr(S)), and finally to the birth of core consciousness (s2), involving a body loop.

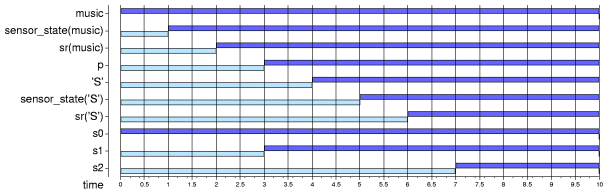


Figure 3: Simulation trace involving a body loop

A similar trace is given in Figure 4, for the case of the as-if body loop. This trace is based on all executable local properties (i.e., LP0 to LP9), except LP3, LP4, and LP5. Again, in all properties, the values (0,0,1,1) have been chosen for the timing parameters e, f, g, and h. As can be seen in Figure 4, in this case the feeling (sr(S)) immediately follows the preparatory state p, without an actual change in body state (S).

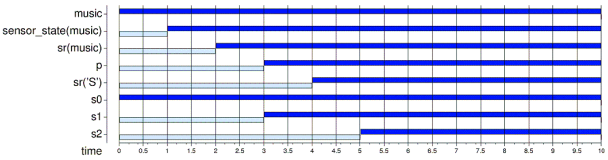


Figure 4: Simulation trace involving an as-if body loop

7 Representational Content

In Damasio's description various types of representation are used, for example, sensory representations and second-order representations. In the literature on Philosophy of Mind a number of approaches to representational content are discussed. In this section three of these approaches are briefly introduced and it is discussed in how far the types of representation used by Damasio indeed can be considered as such according to these approaches.

In (Kim, 1996), pp. 191-192 the *causal/correlational approach* to representational content is explained as follows. Suppose that, some causal chain is connecting an internal state property s and external state property 'horse nearby'. Due to this causal chain, under normal conditions internal state property s of an organism covaries regularly with the presence of a horse: this state property s occurs precisely when a horse is present nearby. Then the occurrence of s has the presence of the

horse as its representational content. Especially for perceptual state properties this may work well.

In (Kim, 1996), pp. 200-202 the concept of *relational specification* of a state property is put forward as an approach to representational content. It is based on a specification of how an internal state property can be related to properties of states distant in space and time. This approach is more liberal than the causal/correlational approach, since it is not restricted to one external state, but allows reference to a whole sequence of states in history.

Finally, the *temporal-interactivist approach* (Bickhard, 1993) relates the occurrence of internal state properties to sets of past and future interaction traces. Thus, like the relational specification approach, this approach allows reference to a whole sequence of states in history (or future). However, whilst in the relational specification approach these states can have any desired type (e.g., internal, external, or interaction states), in the temporal-interactivist approach they are restricted to interaction states (i.e. observations and actions).

In the following sections it is explored whether these approaches can be used to specify the representational content of the relevant mental states that occur in our model (i.e., the states that represent emotion, feeling, and feeling a feeling). The focus is on the causal/correlational approach and the relational specification approach. The temporal-interactivist approach is not discussed. However, the formulae expressing the representational content according to the relational specification approach can be easily translated to the temporal-interactivist approach by replacing the external states that occur in the formulae by interaction states (e.g., replacing music by sensor_state(music)).

7.1 Content of Emotion

Consider the causal chain music - sensor_state(music) - sr(music) - p - S (see Figure 1). Thus, looking backward in time, the external emotional state property S can be considered to (externally) represent the emotional content of the music. On the other hand, the internal emotional state property involved is p. Given the causal chain above the (backward) representational content for both p and S is the presence of this very special music, which could be considered acceptable. However, following the same causal chain, also the state property sr(music) has the same representational content. What is different between p and sr(music)? Why are the emotional responses to the same music different between different individuals? This would not be explainable if in all cases the same representational content is assigned. It might be assumed that state properties such as sr(music) may show changes between different individuals. However, the differences are probably much larger between the ways in which for two different individuals sr(music) is connected to a composite state property p. This subjective aspect is not taken into account in the causal/correlational approach. The content of such an emotional response apparently is more personal than a reference to an objective external factor, so to define this representational content both the external music and the internal personal make up has to be taken into account.

For the relational specification approach the representational content of p can be specified in a manner similar to the causal/correlational approach by 'p occurs if the very special music just occurred', and conversely. However, other, more suitable possibilities are available as well, such as, 'p occurs if the very special music just occurred, and by this organism such music was perceived as sr(music) and for this organism sr(music) leads to p', and conversely. This relational specification involves both the external music and the internal make up of the organism, and hence provides a subjective element in the representational content, in addition to the external reference. This provides an explanation of differences in emotional content of music between individuals.

7.2 Content of Feeling

The representational content of sr(S) according to the causal/correlational approach can consider the causal chain music - sensor_state(music) - sr(music) - p - S - sensor_state(S) - sr(S). Using this chain, sr(S) can be related to both the presence of S, and further back to the presence of the very special music. This steps outside the context of having a reference to one state, which limits the causal/correlational approach. A more suitable approach is the relational specification approach, which allows such temporal relationships to different states in the past; there is the following temporal relation between the occurrence of sr(S), the presence of the S, and the presence of music: 'sr(S) occurs if S just occurred, preceded by the presence of the music', and conversely.

7.3 Content of Feeling a Feeling

The representational content of s0 according to the causal/correlational approach can be taken as the absence of both S and music in the past, via the causal chain: no S and no music - sensor state no S and sensor state no music - no sr(music) and no sr(S) - s0. This can be expressed relationally by referring to one state in the past: 'if no S and no music occur, then later s0 will occur,' and conversely. Formally:

$$\forall t1 \;\; [\, state(\gamma, t1, EW) \models\!\!= \neg S \wedge \neg music \Rightarrow$$
$$\exists t2 \geq t1 \;\; state(\gamma, t2, internal) \models\!\!= s0 \;\;]$$
$$\forall t2 \;\; [state(\gamma, t2, internal) \models\!\!= s0 \;\; \Rightarrow$$
$$\exists t1 \leq t2 \;\; state(\gamma, t1, EW) \models\!\!= \neg S \wedge \neg music]$$

For s1 and s2 the causal/correlational approach does not work very well because these state properties essentially encode (short) histories of states. For example, the representational content of s1 according to causal/ correlational approach can be tried as follows: presence of the music and no S in the past under the condition that at some point in time before that point in time no music occurred. However, this cannot be expressed adequately according to the causal/ correlational approach since it is not one state in the past to which reference is made, but a history given by some temporal sequence. The problem is that no adequate solution is possible, since the internal state properties should in fact be related to sequences of different inputs over time in the past. This is something the causal/correlational approach cannot handle, as reference has to be made to another state at one time point, and it is not possible to refer to histories, i.e., sequences of states over time, in the past. A better option is provided by representational content of s1 as relational specification: 'if no S and no music occur, and later music occurs and still no S occurs, then still later s1 will occur,' and conversely. Formally:

$\forall t1, t2\ [\ t1 \leq t2\ \&\ state(\gamma, t1, EW) \models\!\!= \neg S \wedge \neg music\ \&$
$state(\gamma, t2, EW) \models\!\!= \neg S \wedge music\ \Rightarrow$
$\exists t3 \geq t2\ \ state(\gamma, t3, internal) \models\!\!= s1\ ]$
$\forall t3\ [\ state(\gamma, t3, internal) \models\!\!= s1\ \Rightarrow\ \exists t1, t2\quad t1 \leq t2 \leq t3\ \&$
$state(\gamma, t1, EW) \models\!\!= \neg S \wedge \neg music\ \&$
$state(\gamma, t2, EW) \models\!\!= \neg S \wedge music\ ]$

Similarly, the representational content of s2 as relational specification can be specified as follows: 'if no S and no music occur, and later music occurs and still no S occurs, and later music occurs and S occurs, then still later s2 will occur,' and conversely. Formally:

$\forall t1, t2, t3\ [\ t1 \leq t2 \leq t3\ \&\ state(\gamma, t1, EW) \models\!\!= \neg S \wedge \neg music\ \&$
$state(\gamma, t2, EW) \models\!\!= \neg S \wedge music\ \&$
$state(\gamma, t3, EW) \models\!\!= S \wedge music\ \Rightarrow$
$\exists t4 \geq t3\ \ state(\gamma, t4, internal) \models\!\!= s2\ ]$
$\forall t4\ [\ state(\gamma, t4, internal) \models\!\!= s2\ \Rightarrow$
$\exists t1, t2, t3\quad t1 \leq t2 \leq t3 \leq t4\ \&$
$state(\gamma, t1, EW) \models\!\!= \neg S \wedge \neg music\ \&$
$state(\gamma, t2, EW) \models\!\!= \neg S \wedge music\ \&$
$state(\gamma, t3, EW) \models\!\!= S \wedge music]$

This comes close to the transitions mentioned in Section 5: *the proto-self exists at the inaugural instant - an object comes into sensory representation - the proto-self has become modified by the object*.

The above relational specification is a first-order representation in that it refers to external states of world and body, whereas Damasio's second-order representation refers to internal states (other, first-order, representations) of the proto-self. The relational specification given above only works for body loops, not for 'as if body loops'. A relational specification that comes more close to Damasio's formulation, and also works for 'as if body loops' is the following (**RSP**):

$\forall t1, t2, t3\ [\ t1 \leq t2 \leq t3\ \&$
$state(\gamma, t1, internal) \models\!\!= \neg sr(S) \wedge \neg sr(music)\ \&$
$state(\gamma, t2, internal) \models\!\!= \neg sr(S) \wedge sr(music)\ \&$
$state(\gamma, t3, internal) \models\!\!= sr(S) \wedge sr(music)\ \Rightarrow$
$\exists t4 \geq t3\ \ state(\gamma, t4, internal) \models\!\!= s2\ ]$
$\forall t4\ [\ state(\gamma, t4, internal) \models\!\!= s2\ \Rightarrow$
$\exists t1, t2, t3\quad t1 \leq t2 \leq t3 \leq t4\ \&$
$state(\gamma, t1, internal) \models\!\!= \neg sr(S) \wedge \neg sr(music)\ \&$
$state(\gamma, t2, internal) \models\!\!= \neg sr(S) \wedge sr(music)\ \&$
$state(\gamma, t3, internal) \models\!\!= sr(S) \wedge sr(music)\ ]$

This is a relational specification in terms of other representations (sr(music), sr(S)), and therefore a second-order representation. It has no direct reference to external states anymore. However, indirectly, via the first-order representations sr(music) and sr(S) it has references to external states.

8 Verification

In Sections 3-6, local, executable dynamic properties were addressed, and simulation based on these properties was discussed. In Section 7, dynamic properties to describe representational content of internal states are introduced. These dynamic properties are of a *global* nature. Another example of a more global property is the following:

OP1 music $\bullet\!\!\rightarrow\!\!\!\!\rightarrow$ s2

Informally, this property states that the presence of music eventually leads to the birth of core consciousness (s2). This can be considered as a global property because it describes dynamic of the overall process, whereas the properties presented in Sections 3-6 described basic steps of the process. For both types of global properties (i.e., dynamic property OP1 and the properties specifying representational content), an important issue is *verification*. In other words, are these global properties satisfied by the simulation model described in Sections 3-6? Therefore, the global properties have been formalised, and verification has been applied in two ways: by *automated checks* and by establishing *logical relationships*.

8.1 Automated Checks

In addition to the simulation software described in Section 6, a software environment has been developed that enables to check dynamic properties specified in TTL against simulation traces. This software environment takes a dynamic property and one or more (empirical or simulated) traces as input, and checks whether the dynamic property holds for the traces. Using this environment, the global properties mentioned above have been automatically checked against traces like depicted in Figure 3 and 4. The duration of these checks varied between 0.5 and 1.5 seconds, depending on the complexity of the formula. All these checks turned out to be successful,

which validates (for the given traces at least) our choice for the representational content of the internal state properties. However, note that these checks are only an empirical validation, they are no exhaustive proof as, e.g., model checking is.

8.2 Logical Relationships

A second way of verification is to establish logical relationships between global properties and local properties. This has been performed in a number of cases. For example, to relate OP1 to local properties, intermediate properties were identified in the form of the following milestone properties that split up the process in three phases:

MP1(MtoE) music $\bullet\!\!\rightarrow\!\!\!\!\rightarrow$ sr(music) & sr(music) $\bullet\!\!\rightarrow\!\!\!\!\rightarrow$ S
MP2(EtoF) S $\bullet\!\!\rightarrow\!\!\!\!\rightarrow$ sr(S)
MP3(FtoFF) **RSP** (see Section 7)

For the milestone properties the following relationships hold (for simplicity neglecting 'as if body loops'):

MP1(MtoE) & MP2(EtoF) & MP3(FtoFF) $\Rightarrow$ OP1
LP0 & LP1 & LP2 & LP3 $\Rightarrow$ MP1(MtoE)
LP4 & LP5 $\Rightarrow$ MP2(EtoF)
LP7 & LP8 & LP9 $\Rightarrow$ MP3(FtoFF)

Figure 5 provides the same relationships in the form of a logical AND-tree.

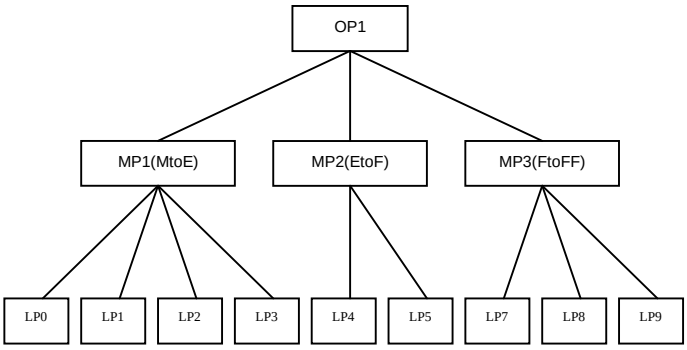


Figure 5: Logical relationships between the dynamic properties

Such logical relationships between properties can be very useful in the analysis of traces. For example, if a given trace that is unsuccessful does not satisfy milestone property MP2, then by a refutation process it can be concluded that the cause can be found in either LP4 or LP5. In other words, either the sensor mechanism fails (LP4), or the sensory representation mechanism fails (LP5).

9 Discussion

The chosen modelling approach describes temporal dependencies in processes at a neurological, not symbolic level. To avoid complexity the model is specified at an abstract level. From the available approaches to representational content from Philosophy of Mind, the causal/correlational approach is not applicable, but Kim's relational specification approach, that allows more complex temporal dependencies, is applicable. Using this approach, claims on representational content made by Damasio have been formalised and supported by means of verification.

Furthermore, an interesting observation that has been made on the basis of the formalisation was that the model predicted the possibility of 'false core consciousness': core consciousness that is attributed to the 'wrong' stimulus. To explain this phenomenon, suppose that two stimuli occur, say x1 and x2, where x2 is subliminal and unnoticed. Then, it could be the case that x2 provokes emotional responses, whilst the conscious feeling that arises is attributed to x1 instead of x2. In terms of our model, this can be simulated by first introducing a subliminal stimulus that yields emotion S (e.g., a cold breeze) followed by the stimulus music. In that case, the conscious feeling would incorrectly be attributed to the music. In personal communication with Antonio Damasio, the existence of this predicted false core consciousness was confirmed.

For the philosophical perspective the paper contributes a case study for representational content which is more down-to-earth than the science fiction style thought experiments, such as the planet Twin Earth, that are common in the literature on Philosophy of Mind, e.g., (Kim, 1996). In addition, the type of representation is more sophisticated than the usual ones essentially addressing sensory representations induced by observing (a snapshot of) a horse or a tomato. Interesting further work in this area is to analyse various arguments given in this literature by applying them to this example.

The analysis approach that is applied in this paper to model Damasio's theory of consciousness, has previously been applied to complex and dynamic cognitive processes other than consciousness, such as the interaction between agent and environment (Bosse, Jonker, and Treur, 2004). In a number of these cases, in addition to simulated traces, also empirical (human) traces have been formally analysed. Using this approach, it is possible to verify global dynamic properties (e.g., specifying the representational content of internal states) in real-world situations.

For recent work in the area of emotion and consciousness, the interested reader is referred to (Prinz

and Chalmers, 2004), Chapter 3, which gives an account for emotions as embodied representations of "core relational themes" such as danger and obstruction.

Acknowledgements

The authors are grateful to Antonio Damasio for his valuable comments upon their questions, and to an anonymous referee for some suggestions for improvements of an earlier version of this paper.

References

Bickhard, M.H., Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence*, 5, 1993, pp. 285-333.

Bosse, T., Jonker, C. M., van der Meij, L., and Treur, J., LEADSTO: a Language and Environment for Analysis of Dynamics by SimulaTiOn. Vrije Universiteit Amsterdam, Department of Artificial Intelligence. Technical Report, 2004.

Bosse, T., Jonker, C.M., and Treur, J., *Representational Content and the Reciprocal Interplay of Agent and Environment* In: Leite, J., Omincini, A., Torroni, P., and Yolum, P. (eds.), Proceedings of the Second International Workshop on Declarative Agent Languages and Technologies, DALT'04, Springer Verlag, 2004, pp. 61-76.

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

Damasio, A., *The Feeling of What Happens: Body, Emotion and the Making of Consciousness.* MIT Press, 2000.

Damasio, A., Grabowski, T.J., Bechara, A., Damasio, H., Ponto, L.L.B., Parvizi, J., and Hichwa, R.D., Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience*, vol. 3, 2000, pp. 1049-1056.

Jonker, C.M. and Treur, J., Compositional Verification of Multi-Agent Systems: a Formal Analysis of Pro-activeness and Reactiveness. *International Journal of Cooperative Information Systems*, vol. 11, 2002, pp. 51-92.

Jonker, C.M. and Treur, J., A Temporal-Interactivist Perspective on the Dynamics of Mental States. *Cognitive Systems Research Journal*, vol. 4, 2003, pp. 137-155.

Jonker, C.M., Treur, J., and Wijngaards, W.C.A., A Temporal Modelling Environment for Internally Grounded Beliefs, Desires and Intentions. *Cognitive Systems Research Journal*, vol. 4, 2003, pp. 191-210.

Kim, J., *Philosophy of Mind*. Westview Press, 1996.

Parvizi, J. and Damasio, A., Consciousness and the brain stem. *Cognition*, vol. 79, 2001, pp. 135-159.

Prinz, J.J. Chalmers, D.J., Gut Reactions: A Perceptual Theory of Emotion (Philosophy of Mind (Hardcover)), Oxford University Press, 2004.

Emergence of Body Image and Dichotomy of Sensory and Motor Activity

Hiroyuki Iizuka Takashi Ikegami

*Department of General Systems Sciences,
The Graduate School of Arts and Sciences, University of Tokyo
3-8-1 Komaba, Tokyo 153-8902, Japan
{ezca,ikeg}@sacral.c.u-tokyo.ac.jp

Abstract

This paper reconsiders a boundary between an agent and its environment. The boundary between a simulated agent and its environment is convoluted in dynamic processes of sensory and motor devices. However the boundary is not a static but a dynamic interface. In order to study the dynamic property of the interface, the un-fixed sensory-motor distinction is introduced. Practically, a mobile arm passively or actively explores an object and attaining some discrimination tasks. We develop the discrimination ability by using a genetic algorithm. A conscious state towards an object is investigated with this framework.

1 Introduction

In order to reorganize the old and new psychological concepts such as ownership, agency and active perception, we need a radical new framework or modeling to supersede sensory-motor flow. Studies on embodied robots and simulations are based on the sensory-motor ideas (Walter, 1950, 1951; Braitenberg, 1984; Pfeifer and Scheier, 1999; Brooks, 1991a,b). For example, Walter discussed cognitive, play-like and social behaviours by synthesizing artificial vehicles. Braitenberg made conceptual robots to discuss the higher functioning of cognition. However, the ideas of ownership and agency are hardly met in this framework.

On the other hand, recent neuropsychological experiments are attacking the problem. Yamamoto and Kitazawa (2001) demonstrated with the arm-crossing experiment that the perceived temporal ordering of haptic stimuli was reversed when the successive stimuli were temporally close enough. Maravita and Iriki (2004) showed that Macaque monkey was trained to use tools to reach food, and showing that its body image was instantly extend to the tip of the tool bar. Ramachandran and Blakeslee (1998) showed that a human body image can be easily created or destroyed by using visual or auditory information. These experiments and others have revealed that body images and ownership have very dynamic nature, which we like to implement in our system.

Our body image and the ownership bridge the gap between the highly abstract sense of "self" and the physical world where our body is situated. Varela (1979) proposed a principle of autonomy stressing the idea of self-generated boundary. Varela exemplified autonomy as a "self" emerged from a chemical system through structural coupling with the environment. In his model, it was shown that some reactive particles created a boundary, which regulated internal reactions of the particles, and maintaining the boundary structure. This circularity of the physical boundary and the internal dynamics produces coherency of self state. The notion of "self" as a dynamic boundary must account for the origin of sensory-motor systems. One such challenge, with respect to a proto-cell system, can be seen in Suzuki and Ikegami (2004).

In this paper, we examine the idea of dynamic boundary in active perception; an agent actively touches an object to discriminate. We assume no explicit distinction between a sensor and a motor. An interface between an agent and its environment is only dynamically constructed. Nevertheless, an agent comes to categorize objects through evolutionary approach. We study the difference between active and passive touch by showing how perception is developed with motion structures.

2 Model

Agents are required to discriminate the number of fans of a windmill by spinning the fans. This task

is inspired by a cookie-cutter experiment by Gibson (1962). Gibson developed a theory of perception based on the active/passive motion structure. In the present task, active or passive pattern is discriminated by the spontaneous motions of the windmill.

The agent has a straight arm which can rotate 180 degrees in a plane (Fig. 1). The arm can spin the windmill by pushing a fan at each impact. The motions of the arm and the windmill are calculated by the following equations;

$$M_a\ddot{\theta_a} + D_a\dot{\theta_a} + F_{arm} + F_{col}(\theta_a, \theta_w) = 0, \quad (1)$$
$$M_w\ddot{\theta_w} + D_w\dot{\theta_w} + F_{col}(\theta_a, \theta_w) = 0, \quad (2)$$

where θ_a and θ_w denote the angles of the arm and the windmill, respectively. F_{col} is a collision term giving a repelling force both to the arm and the fans. The collision is not an instant event but it has a finite width. F_{arm} is a force of the agent to rotate the arm.

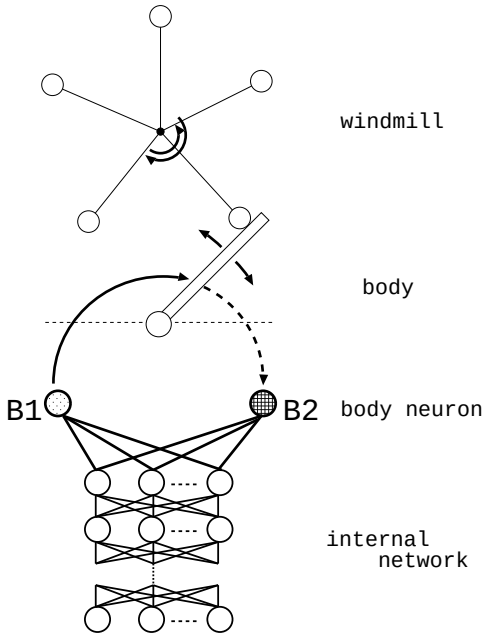


Figure 1: A schematic view of the windmill, the agent and its internal structures. The agent consists of a body with the straight arm whose length is L. The windmill has Q fans. The arm collides with each fan. As the agent's structures, there are body neurons and hierarchical internal neurons (concretely, see Sec. 2.1.). The distance between centers of the arm and of the windmill is determined to be a range $[-k : k]$ where the arm can touch fans. Q and k are assigned according to tasks and L is always set to 100.

2.1 Interface between body and internal dynamics

No explicit sensor-motor flow is pre-fixed so that the distinction between moving and being moved only appears from an internal viewpoint.

Practically, the arm state is assigned to two neurons, which are also connected to other internal neurons. We call these two neurons, 'body neurons.' They are activated or inhibited exclusively in response to the arm state. If one body neuron is more active than the other, one (*actor*) regulates the arm and the other (*observer*) will copy the state of the arm as its state. Each body neuron can potentially play a role of actor and observer depending on their relative activations. By restricting the observation to the state of the arm, we can naturally describe spontaneous moving and being moved by externals in our model, because both motions can be described as changes of arm states. However, the distinction between moving and being-moved becomes implicit. Whether arm motion is moved spontaneously or externally is internally evaluated by investigating the series of activations of the actor and the observer.

By doing this, the fixed sensor-motor relationship is removed. The dynamics of the boundary depends on the two body neurons. Formally, the following equations describe the model:

$$y_{B1} = g_{B1}^{-1}((1-\mu_1)S_1(\theta_a) + \mu_1 g_{B1}(y_{B1})), \quad (3)$$
$$y_{B2} = g_{B2}^{-1}((1-\mu_2)S_2(\theta_a) + \mu_2 g_{B2}(y_{B2})), \quad (4)$$

$$F_{arm} = \alpha\{\mu_1(g_{B1}(y_{B1}) - S_1(\theta_a)) + \mu_2(g_{B2}(y_{B2}) - S_1(\theta_a))\}, \quad (5)$$

$$\text{if} \quad g_{B1}(y_{B1}) - S_1(\theta_a) > g_{B2}(y_{B2}) - S_2(\theta_a), \quad (6)$$
$$\text{then} \quad \mu_1 = 1, \mu_2 = 0, \quad (7)$$
$$\text{else} \quad \mu_1 = 0, \mu_2 = 1, \quad (8)$$

where y_{B1} and y_{B2} are activations of the body neurons. S_1 and S_2 normalize the current arm state, θ_a, from 0 to 1. $g_{B1}(y_{B1})$ and $g_{B2}(y_{B2})$ represent the goal states of the arm, which is a position each body neuron desires to move to if it is the actor. The power of the arm, F_{arm}, is calculated from the difference between the current body state and the goal state of the actor (eq. (5)). The parameters, μ_1 and μ_2, decide which body neuron behaves as an actor or an observer. The body neuron with the larger difference between the goal states and the current state of the arm (eq. (6)) becomes the actor and another becomes the observer. The activations of the actor and the observer are updated as follows (eq. (3) and (4)). In

case of the actor, the goal state of the arm, $g(y)$, is used for the force strength acting on an arm. The goal state will be fed back to the next neural state. In case of the observer, the next neural state will get the current state of the arm, S. Those next neural states are transformed to activations of the body neurons by an inverse function of $g(x)$, which is a transfer function used in a recurrent neural network.

The difference between spontaneous motions and being moved will be detected by a following way. When an arm is moving freely, the activation of the actor precedes that of the observers to the goal state of the actor. The actor and the observer can keep coherency while moving. However, if there is an obstacle or the arm is driven by external forces, the observer's activation can be different in response to the arm state. The coherency is broken at the event. This could be regarded as information flow from the environment to the agent.

The internal dynamics of the agent is controlled by a continuous-time recurrent neural network (CTRNN) (Beer, 1995). The time evolution of the states of neurons are expressed by :

$$\tau_i \dot{y_i} = -y_i + \sum_{j=1}^{M} w_{ji} g_j(y_j), \quad (9)$$

$$g_i(x) = 1/(1 + e^{-x-b_i}), \quad (10)$$

where y is the activation of each neuron, τ is its time constant, b is a bias term, w_{ji} is the strength of the connection from the neuron, j, to i. We adopted a sparse connection among neurons. The neurons are hierarchically organized and the connections of neurons between different layers are only effective and the other connection weights are set to 0 (Fig. 1).

The neurons at the end of layer are connected to the body neurons in the same ways as eq. (9) and (10) (Fig. 1). As an output, the neural network has to choose one between two things in a task. We designed two specific neurons at the opposite side layer of the body neuron's. By comparing the activations of these two neurons, the alternative is represented.

2.2 Tasks : active touching and passive touching

There are four different tasks. In each task, the agent interacts with a 7- or 5-windmill and discriminate it by spinning the windmill (Fig. 1).

In active touch case, an agent can push to spin the windmill. When an arm touches a fan, the windmill rotates clockwise or anticlockwise. On the other hand, when the windmill rotates anticlockwise with a constant speed, an agent cannot control the windmill, and we call it passive touching case.

We use the abbreviations of A7, A5, P7, or P5 corresponding to the task conditions, active or passive, and 7 or 5 fans of the windmill. The parameter, k, is set to $\pi/4$ or $\pi/12$ under active or passive condition, respectively (see Fig. 1).

2.3 Genetic algorithm

Networks were trained by evolving connection weights using a standard evolutionary algorithm. Each artificial genome encodes parameters of the CTRNN: the weights w in [-4,4], time constant τ in [0.4,4], and bias b in [-3:3] as a continuous valued vector. The best agents are carried over to the next generation without any genetic modifications (elitism). Other agents are generated from the best agents by adding a small random values (mutation) from the range [-0.01:0.01]; no crossover is performed. We use 80 agents in this simulation.

Agents' performances are evaluated on the basis of the accuracy of discrimination during the evaluated period of time, which is fixed at 1000 time steps. The fitness value is calculated by multiplying the percentage of correct answers for each task.

3 Results

The best agents after approximately 4000 GA generations can approximately distinguish 7- and 5-windmills, in both active and passive cases. The agents before 2500 GA generations cannot distinguish them. From 2500 GA generations, the fitness is sharply improved as shown in Fig. 2. Each task cannot be achieved one by one at different generations, but all the tasks can be achieved around the same generation.

3.1 Behavior pattern

The best agent basically moves its arm left and right to spin a fan of the windmill. The behaviors under the four different conditions are briefly described as follows. Corresponding to the four conditions, the arm behaviors are given in Fig. 3.

In case of A7, the agent aggressively pushes the windmill when it touches a fan of the windmill, and the windmill rotates clockwise or anticlockwise. After pushing a fan and spinning, a next fan pushes the arm a little bit, and the agent pushing back the fan to the opposite direction.

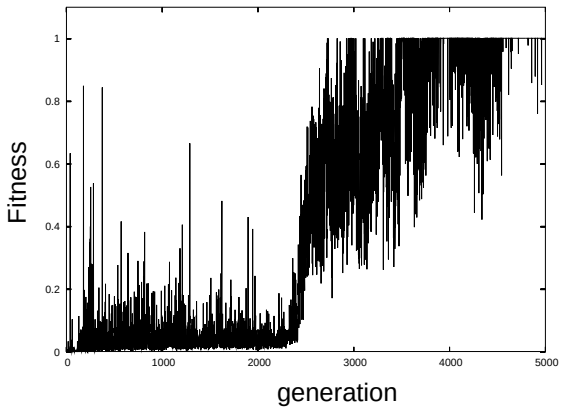


Figure 2: The fitness value of the best agent at each generation.

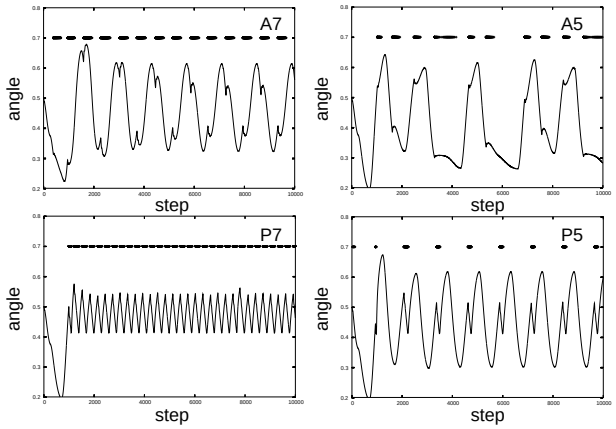


Figure 3: Behaviors of the best agent in its interaction with A7, A5, P7 and P5. In each graph, the time series of the angle of the arm, θ_a, are shown. The dots are plotted at 0.7 when the arm collides a fan.

In case of A5, the agent does not strongly push the fan compared with the condition A7. The arm touches alternately between left and right fans. The alternation periodicity is less frequent than A7 because of sparseness of fans.

In case of P5, in spite of the passive condition, the arm behaves like in case of A7 and A5. Because the number of fans is more sparse than those in P7, it is rare to collide with the fans of the windmill. The agent moves the arm left and right by itself, and the collision timing is regulated by the agent's motion.

Different from those 3 conditions, the arm is pushed into the same direction by the fans under the condition of P7. After being moved, the agent brings the arm back into the center and it is moved again by the next fan. Basically, the agent can tell whether it is 5 or 7 by being moved.

Under each condition, above processes are repeated for discrimination. Figure 4 shows the attractors of the internal dynamics formed at that time. The agent constitutes different attractors, according to the interacting task.

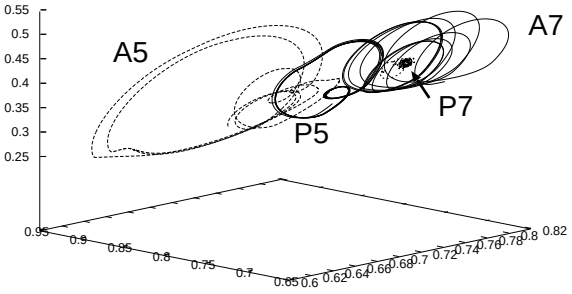


Figure 4: Attractors in the internal dynamics of the best agent in its interaction with A7, A5, P7, and P5. These are plotted by using activations of three internal neurons.

3.2 Dichotomy of sensor and motor

In order to investigate how the interface between the agent and its environment is constructed among different tasks, we perturb the coherency between two body neurons by giving a time delay in the body feedback. In eq. (3) and (4), the function $S(\theta_a)$ is given by the current state of the arm, but we replace it with an arm state several steps before.

Figure 5 shows the agent's performance with the time delay under the condition of A7, A5, P7, and P5. In case of A7, A5, and P5, the agent fails to discriminate depending on the time delay. As explained in the previous section, the task of P5 is basically achieved through the agent's motions although it is a passive condition. The boundary dynamics as an interface under these conditions, where the agent discriminates with active motions, is sensitive to the time delay, which causes a breakdown of the behaviors.

On the other hand, the discrimination can be achieved regardless of the time delay in case of P7. Different from the other three conditions, the coherency of two body neurons is not depending on the timing of being moved. This passiveness can be regarded as a static "sensor" given by the static nature of the boundary dynamics.

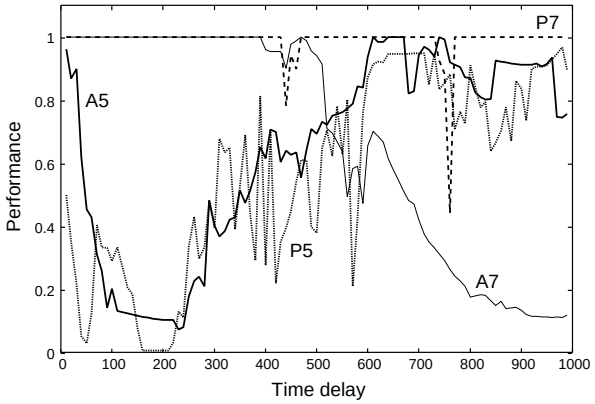


Figure 5: The ratio of the best agent's correctly categorizing under time delay in the body feedback. Horizontal axis means the length of the time delay.

4 Discussions

The ordinary sensory-motor categorization discriminates the sense-data into several domains (Pfeifer and Scheier, 1999). On the other hand, the present paper showed a new way of categorization. As no sense-data is given explicitly, what an agent discriminates is its own motion repertoire that is used to interact with an environment. Body image and ownership is, we believe, cannot be created by the static sense-data. That is why the sensory-motor categorization is not adequate for this matter.

At least two layers are prerequisite for understanding perception or conscious states. One is a physical layer where everything is driven by physical processes and no distinction between operator and operand exists. The other one is a phenomenological layer where everything is described from the first-person's view, based on the notion of "self" which enables subjective distinction between a sensor and a motor.

Two layers are complementary to each other, that is, one side understanding is not enough. A sensor and a motor are equivalent to operator and operand. To define a sensor and a motor first is equivalent to starting a discussion from the phenomenological layer. On the other hand, our present model assumes no distinction between a sensor and a motor apriori at the phenomenological layer. Sensor and motor only emerge with "self" as a result of complex interactions at the physical layer.

Emerging "self" also means an emergence of an interface between "self" and a world. The interface provides how to interact with the world. The world is not just an environment but a world that "self" perceives through the interface, that is, the surplus of signification (Varela, 1992). If we set up a fixed sensor and motor device, an emergence of "self" would not happen. What we want to see is how the subjective distinction such as "moving and being moved" (or tickle and being tickled) emerges and can be sensed. We showed a dichotomy of sensor-like interface from no distinction between a sensor and a motor in this paper. It inevitably requires a link between physical and phenomenological layers. Body image, ownership and active perception are the direct outcomes of this linkage.

Acknowledgements

This work is partially supported by Grant-in aid (No. 15300086) from the Ministry of Education, Science, Sports and Culture, The 21st Century COE (Center of Excellence) program(Research Center for Integrated Science) of the Ministry of Education, Culture, Sports, Science, and Technology, Japan, and the ECAgent project, sponsored by the Future and Emerging Technologies program of the European Community (IST-1940).

References

R. D. Beer. On the dynamics of small continuous-time recurrent neural networks. *Adaptive Behavior*, 3(4):469–509, 1995.

V. Braitenberg. *Vehicles: Experiments in Synthetic Psychology*. Cambridge, MA: MIT Press, 1984.

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

R. A. Brooks. New approaches to robotics. *Science*, 253:1227–1232, 1991b.

J. J. Gibson. Observations on active touch. *Psychological Review*, 69:477–491, 1962.

A. Maravita and A. Iriki. Tools for the body (schema). *Trends in Cognitive Sciences*, 8(2):79–86, 2004.

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

V. S. Ramachandran and S. Blakeslee. *Phantoms in the brain*. New York:Harper Collins, 1998.

K. Suzuki and T. Ikegami. Self-repairing and mobility of simple cell model. In J. Pollack, M. Bedau,

P. Husbands, T. Ikegami, and R. A. Watson, editors, *Artificial Life IX: Proceedings of the 9th International Conference on the Simulation and Synthesis of Living System*, pages 421–426. Cambrige, MA: MIT Press, 2004.

F. J. Varela. *Principles of Biological Autonomy*. Elsevier/North-Holland, New York, 1979.

F. J. Varela. Autopoiesis and a biology of intentionality. In *Proceedings of a workshop on Autopoiesis and Perception*, pages 4–14. 1992.

W. G. Walter. An imitation of life. *Scientific American*, 182(5):42–45, 1950.

W. G. Walter. A machine that learns. *Scientific American*, 185(2):60–63, 1951.

S. Yamamoto and S. Kitazawa. Sensation at the tips of invisible tools. *Nature Neuroscience*, 4:979–980, 2001.

Will and Emotions:
A Machine Model that Shuns Illusions

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

Mercedes Lahnstein
Dept. Of Electrical and Electronic Engineering,
Imperial College , London SW7 2BT
m.lahnstein@imperial.ac.uk

Rabinder Lee
Dept. Of Electrical and Electronic Engineering,
Imperial College , London SW7 2BT
r.lee@imperial.ac.uk

Abstract

Benjamin Libet discovered a neo-cortical 'readiness potential' associated with the spontaneous movement of a finger (Libet et al, 1983). As this happens approximately 350ms before the participant becomes conscious of willing the action, has led Dan Wegner (2002) to develop an illusion-based hypothesis of volition. This paper suggests that the readiness potential is emotional in nature and appropriately unconscious, removing the need to evoke illusions. A machine model is developed which shows how an emotional readiness potential might relate to a legitimate sensation of causation.

1 Introduction

We believe that the problem with Libet's discovery (Libet et al, 1983) of a neo-cortical 'readiness potential' associated with the spontaneous movement of a finger, became controversial because it hinges on a volitional task that does not involve emotion. In general volition does, and it is the objective of this paper to show that if a model incorporating emotion is developed, Libet's Readiness Potential may have an emotional basis. Wegner's (2002) illusion-based hypothesis involves the existence of an unconscious cortical event which both controls a conscious sensation *and* the resulting action which the volitional organism mistakenly interprets as the action being *caused* by the sensation of volition. This, he argues, is akin to believing that traffic lights have changed to green as a result of the will of the observer. Here our addition to Libet and Wegner's reasoning is that choices in an act of volition involve emotional evaluation and this leads to an illusion-free hypothesis. We have developed a mechanistic model which tests a non-illusory hypothesis which still accords with Libet's results and respects a group of basic *synthetic phenomenology* (depictive) rules for conscious mechanisms (Aleksander and Dunmall, 2003 [A&D]).

2 The illusion interpretation

2.1 Libet's data

Libet (1983) wanted to measure the time it took between wanting to do something and doing it. He devised an experiment where the decision to do something had an arbitrariness about it. In line with other similar experiments, he measured the electro-encephalographic recording (eeg) that related to wanting to lift a finger at a time arbitrarily selected by the participant, and the moment of lifting the finger.

Normally this experiment would measure the time of the brain activity when wanting to lift the finger (Bw) and then, the time of both the brain activity that lifts the finger (Bl) and the actual moment of lifting the finger (L). It was known that Bl would occur a small fraction of a second after Bw followed another fraction of a second later by by L. All of this accords with the folk idea that we need brain activity to want something and that a little later this causes other brain activity that activates the muscles that unleash the desired physical action.

To add greater interest to the experiment Libet invented an ingenious way of measuring the moment at which the conscious thought of lifting the finger occurred to the participant. He asked the

participant to observe a dot moving in a circular trajectory on the screen, the trajectory being marked with numbers and to note the number when the conscious thought of lifting the finger occurred. This acted like a clock. Of course the experimenter expected this reading to coincide with the brain activity Bw. The surprise came when it transpired that (Bw) occurred half to three-quarters of a second *before* the participant became conscious of 'wanting'. This brain activity then became known as the *Readiness Potential.*

An obvious interpretation of this result is that the conscious will to do something is not the free event we feel, but it is dependent on an unconscious occurrence in the brain that is initiated in a way that is as yet not properly understood.

2.2 Wegner's interpretation

Dan Wegner (2002), suggested an important philosophical significance of Libet's finding. He proclaimed a 'theory of apparent mental causation':

> *"People experience conscious will when they interpret their own thought as the cause of their action"*

So a totally unconscious neural event causes the wanting and it causes the action a little later. This feels as if the action is caused by the wanting, but the link entirely illusory.

We express a little scepticism on the rush to interpret Libet's results as defining free will as an illusion. First one needs to question exactly what it is that is measured and labelled as Bw. Second, we are able to model at least one mechanism in which the kind of delays encountered by Libet would arise in the course of a non-controversial scheme that links consciousness of a desire to eventual action. The missing feature is the evaluation of emotions associated with available choices. For this we appeal to some 'axiomatic neuromodelling'.

3 Axiomatic neuromodelling

3.1 Axioms: a resumé

While details of our axiomatic approach are fully set out in [A&D], we include a brief resumé here for the sake of completeness. The five axioms are a result of answering the question "What is important to me about my consciousness" and then asking what known informational mechanisms might exist which *are necessary* to sustain such sensations. Sufficiency for these mechanisms is not claimed, but necessity leads to composite systems in which these mechanisms interact to suggest a design for a conscious organism. In this paper they are only used to the extent that we wish to suggest a non-illusory structure that links conscious thought to action. The word 'axiom' is used not in the sense that one might in the starting point of a logical proof, but more in a sense of a set of starting points that are inwardly felt to be fundamental for the design of a model that reflects introspective features of being conscious.

3.1.1 Axiom 1: Being in an out-there-world

What seems a central feature of being conscious is the fact that sensations appear to be situated where they are in the world rather than like pictures or representations in our head. The property required for this is that some neurons not only react to the presence or absence of minimal source events in the world, but they do this conditionally on the position occupied by such an element. The mechanism required for this relies on the presence of sensors that are sensitive to such position as, for example, binocular vision, eye, neck and body movement in vision. As indicated in [A&D] there are neurons in the brain that are selective in this way. This special representation of being in an out-there-world has been called *depictive.*

3.1.2 Axiom 2: Imagining

Referring to visual sensation for a while, it is clear (speaking introspectively) that, if I close my eyes, the visual world does not go away: I can imagine what things look like, that is, what they looked like at some time in the past. The sensation is not quite as vivid as when I am actually looking at something, but there nonetheless.

These 'visions' need not go away when I do open my eyes. Indeed they are part of my visual interaction with the world out there. I sometimes loose my keys and look for them. I form a mental image of what they might look like when I eventually do see them. If I should see a different bunch of keys, the differences between the depiction of these and the mental image are intensely, almost painfully felt. When seeing a well known face, it is known that I can form a sufficiently appropriate mental image of the person even before my fovea has had a chance to look at every feature. That is, the mental image snaps in.

There is another aspect to these inner sensations: they can construct something we may never have seen or experienced as when reading noverls. This is a case where visions are generated by words, but visions could be generated by any of the sensory modalities: the smell of freshly baked bread can trigger scenes from childhood, touching a slimy surface in the dark can create nightmarish visions of unpleasant gutters.

The material implication of these inner visions and memories is that of *feedback* or *re-entry* in depictive neural structures. Having a mental image of something that has happened in the past has a strong material implication: closed information paths in depictive networks must exist which can *sustain* depictive firing patterns. Indeed, axiom 2 could be simply rephrased to say: 'no depiction and no feedback – then no imagination'.

3.1.3 Axiom 3: Attending to input

So far, we have spoken of worlds out there as if the conscious organism just blunders around in them. Nothing is further from the truth. Selecting what we experience in the world and how we think about the world in our imagination, requires some selection mechanisms: attention. The technical detail of how attention is achieved in an external and internal sense is beyond the needs of this paper. Suffice it to say that such mechanisms are largely unconscious. The most telling are eye movements: largely driven by the superior colliculus, they could be determined by the content of the perifovea of the eye (high spatial frequency) by the extrastriate cortex (supply missing meaning) or even the auditory cortex (eyes move in the direction of a sudden noise). But all we feel and experience is the foveal depictive reconstruction of axiom 1.

3.1.4 Axiom 4: Thinking ahead

This axiom and the next are central to this paper, as the paper is directed towards their elaboration and assessment with respect to Libet's data. Thought is not just a process of having static depictions. It is a highly dynamic process. We are constantly thinking ahead, considering alternatives and, every now and then, deciding what to do next. What are the material implications of this possibility? TIn fact, no new machinery needs to be evoked over and above that which we have seen in axiom 2: re-entrant neural networks. It is well known that these are capable of sequence recall as well as the recall of more static experiences.

Again, speaking introspectively, I am looking at a pencil on my desk and deciding that I want to pick it up. This thought is a sensation of my actually doing it in my head, before I do it for real. My depictive areas are producing a kind of film show in my head in anticipation of the real act. This comes from the fact that the depictive areas can learn appropriate depictive sequences as part of the build-up of experience as a sequence of depictive states. That is, as a child I learn to pick things up by trial and error. When I succeed reliably, my visual, tactile and muscular neurons have, together, learned to go from state to state by the same axiom 2 mechanism that allows them to remain stable in one state. There is very little technical difference between learning sequences and learning single stable states. But if there are many possibilities how are these controlled? What is it to want to execute one of the possible plans? This leads to axiom 5 and then to the main meat of this paper.

3.1.5 Axiom 5: Emotions

One of the criticisms levelled at those who speak of conscious machines is that this is one element of humanity that machines cannot have: feelings and emotions. We argue that as these seem to be essential to being a conscious human being they must be essential to being a conscious machine on account of their aid to survival. One should be suspicious of the consciousness of a machine were it not to have mechanisms that play the role of emotions in living organisms.

In the first instance emotions are related to the evaluation of depictive input. Children not more than a few hours old will show signs of fear (facial expression and a retreating action) if a large object moves towards them. The same occurs if the child is allowed to move freely over a glass surface that appears to stretch over a precipice. The child avoids the precipice and shows signs of fear. On the other hand the child shows contentment on being fed when hungry. So, basic emotions such as fear and pleasure, are neural activities that are pre-wired, through evolution, at birth. They have obvious survival value. Others in this innate group are anger, surprise, disgust and distress.

Other emotions and feelings are developed during perceptual life. Feeling hurt after being rebuked or being jealous of the attention someone else is getting are examples of a vast group of such subtle phenomena. On the basis that every scrap of our sensation is due to some neuro-chemical activity, the axiom suggests that such patterns have distinct characteristics that both adapt to be attached to perceptual depictive events as well as imagined events. As planning proceeds according to the mechanisms of axiom 4, predicted states of the world trigger emotional neural firing which determines which plans are preferred for execution and which might lead to unwanted consequences.

This is an area where a great deal of study still needs to be carried out and this paper is an example of such develoment. But one thing is sure, an organism without neural mechanisms for conscious emotional evaluation of thoughts and plans would have its capacity for survival strongly curtailed.

3.2 Comment on the axioms

The main reason for presenting the axioms as sequences going from a felt inner sensation to a generating mechanism that may both be found in the brain and act as a design principle for a conscious machine is to stress an important point. It shows that there need not be any insurmountable gaps in this sequence. If sensation implies mechanism then it at least seems feasible to assume mechanism implies sensation. To seek a science that separates sensation from the action of its material mechanism seems unnecessary.

The axioms should be seen as a necessary set. There is no claim here for sufficiency and the research community is invited to add to the set.

4. An axiomatic 'kernel' structure

Fig. 1 shows a minimal architecture implied by the axiomatic/depictive properties. The perceptual module directly depicts sensory input and can be influenced by bodily input such as pain and hunger. The memory module implements non-perceptual thought for planning and recall of experience. The memory and perceptual modules overlap in awareness as they are both contain locked cells. The emotion module evaluates the 'thoughts' in the memory module and the action module causes the best plan to reach the actions of the organism.

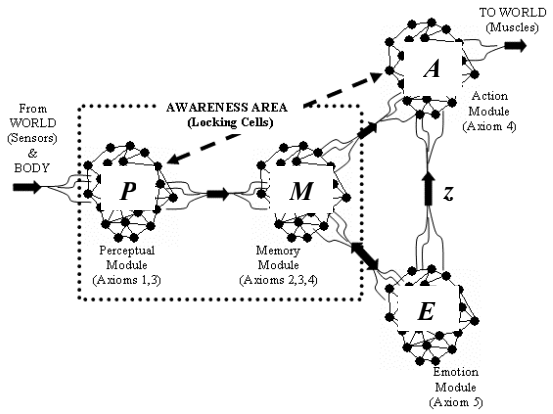


Figure 1. The axiomatic kernel architecture

Each module has its own set of states as follows:

Perceptual states:
P: $\{p_1, p_2, p_3, \ldots\}$

The memory states:
M: $\{m_1, m_2, m_3, \ldots\}$
Where each state is further subdivided into components:

m_j is a triple $(\alpha_j, \rho_j, \varepsilon_j)$ where
α_j is a remembered *action*
ρ_j is a remembered *result of the above action* and
ε_j is a remembered *emotion associated with the above result.*

Emotional states ***E*** are 'wantedness' evaluations that return an intensity ε_j ***(i)*** for state ε_j in ***M.***

Action states $\boldsymbol{A} = \{a_1, a_2, a_3, \ldots\}$ are vectors of *decided* muscular activity that may define the (largely) outward behaviour of the organism.

z is a very important signal: it is generated when the wantedness values for a particular imagined result exceed a threshold ***t*** . This will be clarified in the example below.

5. Emotions and volition in the model – an example

To illustrate the way that this architecture works we imagine a simple scenario of having to make choices: looking at the menu of a pizza restaurant. The menu reveals only three items of food
Pizza; Pasta; Salad.

These, by a process of depiction (ax1) and attention (ax2) become, in turn (say), the states of the perceptual neural module, ***P***. These act as inputs to the memory/imagination module ***M***. Now, it is in this module that the 'thinking' goes on. In this case the thinking has to do with imagining the action that might be taken: eating pizza, pasta or salad. But this is not all one imagines. One recalls the taste of these dishes, and a whole lot of emotions associated with them. This and the operation of the emotion module ***E***, needs some explanation.

5. The emotion process

5.1 Wantedness generation

Emotions are taken to have the following character: first, they are remembered in the context of a predicted action and result of that action. For example, speaking introspectively, in imagining in ***M*** the action of eating pizza, I remember the result of this (the taste) and I also remember the associated emotions. That is, it is the presence of the action state, the taste state and the emotion states that make up the total state of ***M***. There could be several emotion states present at the same time: say, a gustatory pleasure, and also guilt because this is bad for my weight. That is, the predicted result of an action can have a collection of emotions associated with it,

some positive and some negative. The second character of emotions is that they have a value, that is, an intensity, a strength with which an outcome is wanted. Third, under normal conditions, it is possible to resolve unclear combinations of emotions and make decisions in any case.

In the model we are here developing, the role of the ***E*** module is to recognise the imagined emotion and to evaluate it, that is create a value of 'wantedness'. This is not an instantaneous process and may involve attending to the available actions several times. The feeling is a familiar one: it takes a while to ponder the content of a menu and the actions and their results are visited several times. Whenever a choice of action is visited in ***M*****,** the total of all the related emotional values is summed up to give a 'how much wanted' value for each action.

So, in our emotional model, ***E*** generates a signal z when a sufficient level of wantedness is achieved. This is fed back to the imagination/memory module and holds that module in the action that has given rise to this high level. At the same time this signal is also sent to the action module enabling it to be set up to drive muscles as required by the action that is currently in ***M*** that the organism has decided to take. Of course this scheme allows for the ***E*** module never to reach an adequate 'wanting' intensity. To deal with this and conflicting or unclear decisions implies that in ***E*** there should be a random process which modulates that computed value of z. The object of the random process is to provide an arbitrary enhancement of the 'wantedness' score in such a way that higher scores are more likely to trigger the action, but that lower score can also achieve this but with a lower likelihood. Even where there is no wantedness, just through the need to make an arbitrary decision (as we shall see in Libet's experiments) the random process can make the arbitrary decision of when to act.

5.2 The Random Process

Neural networks are very good at random processes, and it is often assumed that the neural structures of the brain are capable of this. For example, in our simulation of the example discussed here, we used a very small 10-neuron net which, due to lack of training, was simply producing binary states. The 'score' of the net was the count of neurons at 1 at any one time. This has normal distribution characteristics that are,

Probability of scores 0 and 10 = .097%,
1 and 9 = 0.97%, 8 and 2 = 4. 4% , 7 and 3 = 11.7%, 6 and 4 = 20.5%, 5 = 24.6%.

Now we can go back to the example and list the emotional value that might be in force at one particular moment that our organism is trying to decide what to eat. This can be set out as table 1 :

Table 1: Total wantedness values for three dishes

Menu item	Pleasure	Lack of Guilt	Total
Pizza	3	0	3
Pasta	2	0	2
Salad	1	1	2

Now say that to make a decision (generate signal z) the ***E*** machinery has to equal or exceed a value of 7 (chosen arbitrarily for illustration) when adding the total emotional value to that of the random process net. So, the decision for pizza will be made when the ***M*** module is in the Pizza state and the random process generates 4 or more. This is obtained by adding the probabilities of random values from 4 to 10 all of which will cause the action. This sum turns out to be 82.8%. Similarly, when ***M*** is "thinking" of Pasta or Salad the random process must generate a value of at least 5. Summing the probabilities of generating 5 or greater indicates that the decision to eat the dish currently thought of will be taken in 62.3% cases.

A way of interpreting these results is to think of 100 (99 to be precise to have a number divisible by 3) people in this restaurant. They all have exactly the same feelings about the three dishes and each has the same random process. Each will be in one of the three food states, say about a third (33) in each. So of the group thinking 'pizza' (0.828x99/3) will make their decision to eat pizza that is, about 28 will make a firm decision to eat pizza. We note that 5 will make no decision at all, and return to consider the next item on the menu. Similarly, of the other two groups of 33 about 20 will choose pasta in the pasta group and 20 will choose salad in the salad group.

Of course these figures depend on the threshold of 7 that has been chosen to do the calculations. Had a higher threshold been chosen, the restaurant clients would be seen as being far less prepared to make decisions. So this threshold can be seen as a sort of 'mood' emotion. A hungry, relaxed mood (low threshold) will lead to a quick decision whereas, an anxious, picky mood (high threshold) will cause the organism to agonise longer before taking a decision.

The point of all this is to show that uncertainties and conflicting emotional values *could* be represented in machinery as shown in fig.1. We stress that it is not the case that this precise machinery is thought to exist in human brains. It is more an expression of the axiomatic stance which suggests that an architecture might at least be envisaged which appears to have the characteristics of emotional

evaluations that go on in the head when one is trying to make a decision. But this needs a bit more discussion.

5.3 How does it feel?

The main tenet of the axiomatic approach is that only that which is depicted is experienced as a meaningful sensation. In figure 1 only the ***P*** and the ***M*** modules are depictive, ***P***, directly and ***M*** indirectly as a memory of the states in ***P***. Therefore it is quite true to say that we are fully conscious of what our options are and the nature of the emotion that comes with them. We have argued earlier that some such feelings are 'wired in' as, for example, fear, pain and pleasure are internally generated neural signals the reach the depictive areas in order to come into consciousness as 'visceral' sensations.

What is new here is that we have imagined a little further than in the bland statement of the axioms how the emotion module could use a random process to control the generation of action. This in situations where a direct reactive response (such as swiping my hand at a fly that has settled on my nose) is not possible due to there being several choices. While this process is not in the depictive part of the mechanism and one would not be conscious of it, the generation of z comes into consciousness as it holds the state of ***M*** for long enough to transfer the imagined action to the action module. This 'freezing' is what we would describe as the moment of consciousness that a particular action would be taken. This now puts us into a position where we can re-visit Libet's findings, and decide that perhaps will is not an illusion after all.

6. Lifting Libet's finger

In Libet's experimental setting, the decision to be made is not 'what' but 'when'. While there may be thoughts and emotions present about what one is meant to do, they do not have a direct bearing on the process. We suggest that the only thing that remains at work is the random process. Everything in the ***M*** module is set up to lift the finger, that is the intention is depicted, but the random 'wanting' machinery is on its own as there are no emotions to evaluate. In such a situation the threshold for generating z should be within the range of values produced by the random process on it's own. For the sake of an explanation, we assume that the same random process is at work and this means that the threshold 7 will be reached if the random process produces z with the sum of the probabilities of generating 7, 8, 9 and 10. Just out of interest this turns out to be 17.2%, that is, somewhat lower than the 'choice' decisions in which emotions are evaluated.

It is now possible to create a hypothesis that makes the unconscious generation of the Readiness Potential less mysterious and will less of an illusion. First, we submit that the generation of something like z in ***E*** *is* the readiness potential. But this is not the source of the willed action, just an emotion-like trigger that the action should take place. So it is hardly surprising that this trigger should be generated before the depiction of the action in ***M*** freezes, which is the moment at which the participant would look at the clock.

In other words the sequence of events goes like this: the desire to lift the finger at some point is fully depicted in ***M***, but without input from ***E***. This activates the random process which with some delay determined by the probability (i.e. not the same each time) of exceeding the threshold, generates z (the readiness potential) to which ***M*** reacts, say, half a second later and the action is transferred to the muscles a little while after that. But there seems little doubt to me that it is the initial desire depicted in ***M*** that is the cause of these events. So, it now becomes possible to summarise a perspective on the concept of will and all it entails.

7. Free will: a summary

It is possible to point at a folk theory of volition: we visualise something we desire – we act to get it. But our cultural inheritance from philosophy and religion and some recent neurological measurements would not leave it at that. Important questions have to be answered, and here we attempt to summarise what might have been gleaned through the depictive/axiomatic approach.

The first of the important questions is that of *freedom*. In what sense can the system in fig. 1 convey the notion of freedom? More pertinently, say that some elaborate form of this system were present in my brain, how is it that it makes me *feel* free to make my decisions in an unfettered way? We suggest that this freedom is felt at least at two levels. Back in the pizza restaurant scenario, (once again, using introspective language) I know, as I can know anything else, that in going to the restaurant, I shall be offered a choice of food from which I will be able to exercise my power of choice, with their emotional overtones and all. This is knowledge like any other: like that when I go into my study I know that will find a computer on the desk or like that when I go to Venice I will have to leave my car in a garage outside the watery city. All this is due to the natural mechanistic implications of axiom 2: areas such as ***M*** provide access to knowledge and experience, and my knowledge of restaurants tells me that I will be able to make a choice that suits me at the time, dependent on my moods and emotions.

Of course, someone could say that it is all predetermined. But it does not *feel* that way just because I am aware of having made different choices under similar conditions. So, predetermined or not, the feeling that what I will choose will be best for me at the time, is good enough not to feel constrained. What would *not* feel free would be the prediction from my prison cell that the same slop as always will arrive at midday.

The above is the first, higher, level of feeling the freedom of will. The second, lower, level, is the mechanism of evaluating emotional states for a series of attentional phases directed (say) at the restaurant menu. The cycling and eventual freezing of the state all occur in ***M***, which, according to the depictive axioms, is felt by the organism. Given language and choice of the most wanted item the organism would describe this as "I felt free to look at the choices offered on the menu and chose the one that appealed to me most". Or if a less wanted item was chosen this might be described as "I chose salad despite the fact that I don't like it all that much, but I know is good for me." In the finger lifting exercise most of us would admit that we didn't know what made us lift a finger at a particular time and that that moment seemed arbitrarily chosen.

The second question about what factors influence my choice is answered by the way we have suggested that emotional evaluations work. Emotions are recalled in ***M*** and the factors of 'wantedness' are computed in ***E***. Of course this process has not been properly elaborated here and is the subject of current research. Open questions relate to how the evaluations get developed through learning and how 'thresholds' develop and change with moods. Finally, the mechanisms described here clarify the involvement of axiom 4 and 5 as the basis of free will. Axiom 4 is the cycling in ***M*** and axiom 5 is the operation of the ***E*** machinery.

7. Will: a philosophical coda

While engineering arguments have been heavily employed in the last few paragraphs above, philosophers may not be happy with this. We conclude this chapter by setting out the logic of the argument in a series of assertions.

For an organism with a 'brain' to have a sensation of free will

1. There exist areas of the brain that support consciousness through having the depictive property (axioms 1 & 2)
2. There exists in the brain a depictive mechanism for cycling through the choices prompted by a perceived external event or an internal imagined event.
3. Cycling through the states in 2 includes memories of emotions associated with the choice states.
4. There exists in the brain a non-depictive evaluational mechanism (non-conscious therefore) that accumulates 'wantedness' values for the emotions associated with each choice. When wantedness exceeds some threshold the current choice state is translated into action.
5. As part of the evaluational mechanism, there exists a random process which adds to the wantedness values helping to resolve situations of conflict or lack of emotional value.
6. Through the depiction of the freezing of the cycling mechanism due to a wantedness trigger, the organism feels that actions are taken among choices according to how much something is wanted.

Conditions 1 and 2 are fundamental postulates which, if denied, block proceeding with the rest. Denying 3 requires a denial that emotions are involved in making choices. Such a denial would contradict common experience and documented material such as Damasio (1995). 4 and 5 are then the basis of the main hypothesis presented here, their denial or confirmation is a matter for both neurological and modelling research.

Acknowledgements

I. Aleksander wishes to thank the Leverhulme Trust for having provided an Emeritus Fellowship that makes involvement with this work possible. Rabinder Lee is supported by an EPSRC grant.

References

Benjamin Libet, and E.W.Wright, and D. K. Pearl,'Time of conscious intention to act in relation to onset of cerebral activity (readiness potential): The unconscious initiation of a free voluntary act', *Brain* , **106**, pp 623 – 42, 1983

Dan. M. Wegner. *The Illusion of Conscious Will*, Cambridge MA: MIT Press, 2002

Igor Aleksander, and Barry Dunmall: 'Axioms and Tests for the Presence of Minimal Consciousness in Agents' *Journal of Consciousness Studies*. **10**, pp 7-18, 2003

Antonio Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain,* London: Picador, 1995